Environmental Determinants of Lexical Processing Effort

Scott McDonald

PhD
University of Edinburgh
2000

Declaration

I declare that this thesis has been composed by myself and that the research reported here has been conducted by myself unless otherwise indicated.

Scott McDonald
Edinburgh, May 4, 2000

Acknowledgments

I would like to thank my supervisors, Richard Shillcock and Chris Brew, for their advice, help and encouragement over the last three years. The thesis also benefited from discussion with and contributions in various forms from Will Lowe, Frank Keller, Mirella Lapata, Mike Ramscar, Martin Corley, Simone Teufel, Louise Kelly, Padraic Monaghan and Ellen Bard.

I would also like to thank Betty Hughes, Dyane McGavin and the Computing Support Team at the Centre for Cognitive Science.

For financial support, I am grateful to the National Sciences and Engineering Research Council of Canada, the Overseas Research Student Awards Scheme, the Sir Ernest Cassels Education Trust, and the Institute for Adaptive and Neural Computation.

Abstract

A central concern of psycholinguistic research is explaining the relative ease or difficulty involved in processing *words*. In this thesis, we explore the connection between lexical processing effort and measurable properties of the linguistic environment. Distributional information (information about a word's contexts of use) is easily extracted from large language corpora in the form of co-occurrence statistics. We claim that such simple distributional statistics can form the basis of a parsimonious model of lexical processing effort.

Adopting the purposive style of explanation advocated by the recent rational analysis approach to understanding cognition, we propose that the primary function of the human language processor is to recover meaning from an utterance. We assume that for this task to be efficient, a useful processing strategy is to use prior knowledge in order to build expectations about the meaning of upcoming words. Processing effort can then be seen as reflecting the difference between 'expected' meaning and 'actual' meaning. Applying the tools of information theory to lexical representations constructed from simple distributional statistics, we show how this quantity can be estimated as the amount of information conveyed by a word about its contexts of use.

The hypothesis that properties of the linguistic environment are relevant to lexical processing effort is evaluated against a wide range of empirical data, including both new experimental studies and computational reanalyses of published behavioural data. Phenomena accounted for using the current approach include: both single-word and multiple-word lexical priming, isolated word recognition, the effect of contextual constraint on eye movements during reading, sentence and 'feature' priming, and picture naming performance by Alzheimer's patients.

Besides explaining a broad range of empirical findings, our model provides an integrated account of both context-dependent and context-independent processing behaviour, offers an objective alternative to the influential spreading activation model of contextual facilitation, and invites reinterpretation of a number of controversial issues in the literature, such as the word frequency effect and the need for distinct mechanisms to explain semantic and associative priming.

We conclude by emphasising the important role of distributional information in explanations of lexical processing effort, and suggest that environmental factors in general should given a more prominent place in theories of human language processing.

Contents

1.
Introduction

> *Words are strange things. They are obviously essential, yet insufficient as ingredients of verbal messages. They are apparently easily identified by laymen, in spite of the fact that linguists prefer to leave them undefined. This does not at all seem to affect their utility as stimulus and response entities whose 'meanings' we may try to assess in the psychological laboratory…*
>
> — R. Rommetveit (1968, p. 97)

This thesis is about the relevance of simple distributional statistics to human language processing. More specifically, it attempts to establish a connection between measurable properties of the linguistic environment and the effort involved in processing words. In this introductory chapter, we outline the fundamental assumptions behind the project.

1.1 Understanding processing effort

One of the most important, yet perhaps undervalued products of over a century of psycholinguistic research is a set of sophisticated techniques for measuring the effort involved in processing language. Methods of measuring processing effort in the laboratory range from the obvious to the ingenious. Reaction times, whether measured using motor responses (such as pressing a button) or vocal triggers (such as pronunciation of a letter string), have proved immensely useful and are probably the most utilised type of measuring instrument. Less obtrusive techniques, such as

the monitoring of eye movements during natural reading, are becoming increasingly popular, because they eliminate the need for subjects to perform a task peripheral to the behaviour of interest.

The fact that processing effort (or ease) is *measurable* has allowed much to be inferred about the nature of the cognitive processes and representations underlying language comprehension and production. Until recently, measurements of processing difficulty were used almost exclusively in the pursuit of *mechanistic* explanations of language behaviour. Researchers have attempted to unravel the complex cognitive machinery involved in, for example, the recognition of a string of alphabetic characters as a real word. Experiments with human subjects are typically carried out in order to test and refine computational models of the proposed mechanism. This approach has provided a stimulating and productive research agenda, and has resulted in a deeper understanding of the architectures and algorithms that are fundamental to cognition.

Because measurements of processing difficulty are simply numerical quantities – no assumptions about the cognitive mechanisms which actually instantiate the process are required – they can also be used to evaluate purely *purposive* explanations of human language behaviour. The purposive style of explanation – characterising a process in terms of its ultimate function – forms the basis of the recent *rational analysis* approach to understanding cognition (see Chater & Oaksford, 1999, for an overview). In this approach, the goals of the task (or the problem to be solved) are first identified, and it is assumed that the cognitive system is adapted to optimally address the goals. A function is then derived which optimally relates the processing goals to a formal model of the environment, taking into account reasonable computational limitations. Thus, the rational analysis approach aims to explain a cognitive process at a high level, while making minimal assumptions about the actual mechanisms involved in implementing the process. Both the behavioural phenomena and the mechanisms responsible are considered to arise from the interaction of the goals of the cognitive system with the environment (Anderson, 1990).

An important premise of rational analysis is that "… many aspects of cognition can be viewed as optimised (to some approximation) to the structure of the environment." (Oaksford & Chater, 1998, p. 3). A cognitive function can be considered to be *optimal* if it can be shown to minimise the cost incurred in carrying out a particular task. This view – that human cognitive behaviour is optimally

adapted to its environment – parallels the standpoints taken in fields such as animal behaviour and evolutionary psychology.

In this thesis we adopt several key features of the rational analysis approach to cognition,[1] and apply them to explaining *lexical* processing effort – the relative ease or difficulty involved in processing *words*, both in and out of context. The central aim of this thesis is to contribute towards an understanding of lexical processing effort, but from a previously unexplored perspective – the 'computational level' of explanation (Marr, 1982). In forming a computational-level theory or model, one makes hypotheses about the function of the cognitive process (what the process should compute), and then collects experimental data in order to test and constrain the model. The success of the model is measured principally through its ability to capture detailed behavioural data. In order to provide a computational-level explanation for lexical processing effort, we shall attempt to satisfy three subgoals: (a) to characterise the primary function of the lexical processor, (b) to develop a computational model of the cost incurred by the processor in carrying out this function, and (c) to test the predictions of the model using empirical data – measurements of lexical processing effort.

Lexical processing effort is widely held to be sensitive to perceptual factors (*eg*. word length in letters or phonemes, typographic case, clarity), lexical/semantic variables (*eg*. grammatical category, familiarity, corpus frequency, concreteness, ambiguity), and contextual influences (from the syntactic, semantic and pragmatic context). Of course, we do not pretend to be able to account for all of these influences on processing effort; the bulk of this thesis will be concerned with the effects of 'semantic' variables and capturing the influence of semantic context. Under the conventional mechanistic approach, a complex explanation or set of explanations would appear to be required in order to account for the variety of factors affecting lexical processing outlined above. But do they really need to be so complex? We will see that one advantage of the purposive style of explanation is that such apparent complexity can be reduced.

[1] Although closely related, our approach differs from rational analysis regarding the optimisation of lexical processing to the structure of the environment. Although we consider the problem faced by the language processor to be the efficient recovery of meaning, it is not yet clear how an *optimal* solution to this goal could be computed.

1.1.1 Building semantic expectations

The first step towards understanding lexical processing effort is to identify the purpose of the processing task. It is uncontroversial that the primary function of human language comprehension is the recovery of *meaning* from an utterance. We propose that a 'local' goal – relevant to the general goal of recovering the meaning of an entire utterance – is the recovery of the meanings of individual words in the utterance. It is satisfaction of this local goal that gives rise to variability in lexical processing effort.

We assume that *efficiency* is an adaptive property of the human lexical processor. A function or process is maximally efficient when it makes the most effective use of available resources in order to accomplish the task at hand. Under this definition, the lexical processor is maximally efficient if it can draw upon available resources in order to minimise the cost of recovering word meaning. Efficiency refers to a net effect: adopting a strategy that reduces the effort involved in recovering meaning in certain circumstances will often be at the expense of increasing the effort expended in others.

The cost of recovering word meaning will be minimised if the processor is able to exploit a source of prior knowledge about the meaning of the word in question. One obvious source for this prior knowledge is the semantic information available from the preceding context. (We return to this point in section 1.1.3 below.)

We propose that a useful strategy for maximising efficiency is to exploit prior knowledge in order to build *expectations* about the meaning of upcoming words. If the *meaning* of an upcoming word is predictable (in some sense), then processing of a *word* that conveys that meaning should be easier than if no expectations about its meaning can be formed. An expectation-building strategy would therefore increase efficiency by reducing *uncertainty*. Thus, we see the development of semantic expectations as a reasonable strategy for efficiently recovering the meaning of the words comprising an utterance, which would have the effect of optimising the more general goal of extracting the meaning of the entire utterance.

Besides efficiency, other advantages to the hypothesised expectation-building strategy include (a) robustness to noise, and (b) the ability to cope with ambiguity. When the linguistic input is *noisy* (due to factors such as a high level of background noise, poor acoustics, or speaker disfluency), the ability to predict the meaning of upcoming words is a desirable property. Another way of looking at this is in terms of *redundancy*: it is intuitive that an utterance contains a great deal of semantic

redundancy (*ie*. the meaning of what has already been said is informative about the meaning of what is still to come), and that people are able to exploit redundancy to aid word identification and consequently improve their overall chance of communicative success (*eg*. Miller, Heise & Lichten, 1951). If the input is partly obscured by noise or otherwise incomplete, then semantic expectations could help to 'fill in' whatever is missing.

Lexical ambiguity is a prevalent, natural feature of all human languages. Yet it is something that people deal with effectively and almost entirely without awareness that any ambiguity is present. Again, the formation of semantic expectations would be a useful strategy for narrowing down the range of possible meanings invoked by an ambiguous word. In order to cope with ambiguity, it would be advantageous if the 'semantic space' could be constrained to those aspects of meaning that are relevant to the communicative context. It seems clear that any processing strategy that succeeds in minimising the potentially disruptive effects of lexical ambiguity would automatically increase efficiency.

What are the anticipated *disadvantages* to the proposed expectation-building strategy? We might find that encountering a word whose meaning is completely unexpected *increases* processing effort (*ie*. reducing efficiency). It is clear that there must be a trade-off – if semantic expectations are too detailed, then the disruption caused by an unexpected upcoming word would counteract the benefits obtained in normal circumstances. In order to maximise efficiency, an adaptive processor should build expectations about meaning at an appropriate level of precision to give rise to an overall reduction in lexical processing effort.

The hypothesis that the comprehension of natural language involves an automatic process of expectation-building is by no means new, although previous research has typically focused on providing mechanistic accounts of the process.[2] For instance, Tanenhaus and Lucas (1987) proposed that the representation of the current discourse developed by the listener, together with the local context and world knowledge "… can often be used to generate reasonable expectations about the incoming speech" (p. 224). They suggest that this strategy could increase lexical

[2] Although making no claims about psychological reality, Lauer (1995) argues that semantic expectations should play a central role in natural language processing (NLP). In his 'meaning distributions' theory, semantic expectations are represented as a probability distribution over the space of possible meanings. Using this approach, Lauer constructed a probabilistic model for the task of paraphrasing English noun compounds.

processing efficiency in context if the structure of the mental lexicon was similar to the structure of the environment.

Becker (1980) put forward an 'expectancy-based' account of lexical processing effort in his explanation of the single-word priming effect. Under this view, people are assumed to use a prime word in order to *consciously* generate a set of candidate related words. If the upcoming word is a member of this 'expectancy set', then processing is facilitated; if the target word is not found in the expectancy set, then inhibition is the result.

In summary, the formation of expectations has been proposed here and in previous research (in various forms) as the cognitive strategy responsible for reducing processing effort in certain circumstances. What distinguishes the contribution of this thesis from previous work is that we present a practical methodology for *quantifying* these expectations.[3]

1.1.2 Operationalising word meaning

In order to model the process of building expectations about word meaning, some means of *representing* meaning is required. This is no longer a major obstacle. Useful objective methods for representing word meaning have been developed over the last few years; these data-intensive techniques use simple distributional statistics collected from large language corpora to represent a word as its distributional pattern of use (*eg.* Lund & Burgess, 1996; Redington, Chater & Finch, 1998). In this type of approach, a word is represented in terms of its relationships with other words, where 'relationship' is defined as simple co-occurrence (*eg.* within a sentence or a 'window' of words).

Representing meaning in this way allows formalisation of a critical component of the proposed expectation-building strategy for the efficient recovery of word meaning: the nature of semantic expectations and the nature of the meaning of the words to which they are to be compared. A formal representational medium permits the high-level concept of building semantic expectations to be implemented in a computational model; this is the main objective of Chapter 6.

The corpus-based approach to representing meaning is also attractive for modelling the *development* of word meaning. If simple distributional information

[3] In fact, we will investigate two related methods (derived from the same corpus data); one viewed from the representational perspective, and the other in incremental processing terms.

really does form the basis of a word's cognitive representation, this implies that the processor is sensitive to the structure of the environment during language development. As experience with a word accumulates, more information about its contexts of use becomes encoded, with a corresponding increase in the ability of the language learner to use the word accurately and appropriately. Although certainly interesting, the developmental direction is not pursued here (for further discussion, see Lowe, 1997).

Of course, one cannot claim that a word's co-occurrence pattern relates to its meaning in some transparent way without empirical justification. Our assumption that lexical representations derived from distributional statistics are useful for operationalising word meaning is verified against behavioural data. Validation is done indirectly; co-occurrence patterns are assumed to contain *semantic* information because a measure of the similarity of two words' distributional representations both corresponds to empirical measures of semantic similarity and predicts semantic priming effects. The first part of this thesis is concerned with the methodology behind the construction of corpus-based representations of word meaning and assessing their psychological validity.

Finally, a practical advantage to representing a word as its distributional pattern of use is the ease by which a probabilistic interpretation can be imposed on the data. This will prove useful in Chapter 6 when we develop a Bayesian method for revising semantic expectations on the basis of contextual information.

1.1.3 Context-independent and context-dependent processing

A central concern of this thesis is explaining the influence of *context* on lexical processing effort. The majority of psycholinguistic research on lexical processing has restricted investigation to the processing of isolated words, under the assumption that a comprehensive and carefully worked-out account of lexical processing in the absence of context will transfer easily to an explanation of context-dependent processing behaviour. Unfortunately, this is not necessarily the case, as robust effects of lexical variables on the recognition of an isolated word often disappear when the same word is embedded in a connected linguistic context (*eg*. Schwanenflugel & Shoben, 1983).

We argue that a general computational-level model should account for variability in lexical processing effort in both the *absence* and *presence* of context. Although it is intuitive that expectations about the meaning of upcoming words could be formed

from the preceding context, it is not immediately obvious how the same strategy would apply to the processing of an isolated word. To model the effort involved in recovering the meaning of an isolated (or utterance-initial) word, we simply assume that the source of prior knowledge on which semantic expectations are based is *uninformative*. This is straightforward to represent probabilistically, and fits naturally into our general framework. This integration of context-independent and context-dependent lexical processing behaviour into a single account is one of the most interesting and important contributions of the thesis.

Applying the expectation-building model to the processing of isolated words illuminates the distributional differences *between* words: words vary according to the complexity (or restrictiveness) of the contexts they occur in. We call this lexical property *contextual distinctiveness*, and demonstrate that, besides being predictive of processing effort, contextual distinctiveness (CD) correlates with subjectively-measured lexical properties such as semantic ambiguity.

The nature of a word's cognitive representation is thus assumed to influence the effort involved in processing that word in isolation. CD directly instantiates the hypothesis that the features of a word's environment form an integral part of its representation, and provides a novel technique for analysing the information contained in a *single* corpus-based lexical representation. Although existing models derived from distributional statistics (*eg*. Landauer & Dumais, 1997; Lund, Burgess & Atchley, 1995) allow words to be easily *compared* (*eg*. by using any of a number of measures of distributional similarity), these representational models have little to say about differences in processing effort *between* words. This is a serious limitation for this type of model, which this thesis hopes to rectify.

1.1.4 The main hypothesis

Measuring the cost or effort of human lexical processing can be done using a variety of techniques. Up to this point, we have refrained from describing how this cost could be computationally modelled.

Our working hypothesis is that the processor adopts the strategy of building expectations in order to reduce uncertainty about the meaning of upcoming words. In order to estimate this quantity, a continuous measure is required that reflects how closely the meaning of an upcoming word matches expectations. By representing words as their distributional patterns of use, we have a practical method for operationalising meaning. This allows us to compute the amount of *information*

conveyed by a word about its meaning as, roughly speaking, the *difference* between the *expected* meaning and the *actual* meaning. So, the more precise the expectations, the closer the match between prediction and reality, and the less information conveyed. We can now state the main hypothesis: the amount of information conveyed by a word about its meaning is predictive of the processing cost incurred by that word:

$$E(w) \propto I(w)$$

More explicitly, this expression states that the effort of processing word w, $E(w)$, varies directly with the quantity of information w conveys about its meaning, $I(w)$. This quantity is easily calculated using the tools of information theory, full details of which are provided in Chapter 5.

The bulk of this thesis is concerned with testing this fundamental principle. We apply the ideas introduced here to a wide range of language processing phenomena, with the aim of providing a parsimonious, computational-level explanation for a disparate set of behavioural data.

1.1.5 Summary

The above constitutes a fairly intricate argument; it is worthwhile distilling it down into its essential points. The following summarises the main thrust of the argument:

- The problem faced by the human language processor is conceived as the recovery of meaning from an utterance.
- The processor is adapted to solve this problem in a way that minimises processing effort (*ie.* maximises efficiency).
- An effective strategy for maximising efficiency is for the processor to use prior information to build *semantic expectations* about the meaning of upcoming words.
- The distributional information contained in large language corpora supplies a useful medium for the objective and psychologically plausible representation of word meaning.
- The difference between the expected meaning of a word and its 'actual' meaning can be expressed as the quantity of *information* conveyed by the word about its contexts of use.

- This information-theoretic quantity can be easily computed for a word occurring in isolation or in context; for isolated words, semantic expectations are assumed to be relatively uninformative.
- The central hypothesis to be tested is that this information-theoretic measure is predictive of behavioural measurements of lexical processing effort.

1.2 Overview of the thesis

The empirical work presented here (new experiments, simulations and reanalyses) addresses a broad range of psycholinguistic phenomena; therefore we shall include the relevant literature reviews in the appropriate sections rather than presenting them in a single chapter.

The first part of the thesis (Chapters 2-4) is concerned with introducing the basic methodology and evaluating the psychological relevance of the corpus-based approach to representing word meaning.

In Chapter 2, we briefly discuss the motivation for creating lexico-semantic representations[4] from simple distributional statistics, and describe the techniques involved in their construction. Chapter 2 also delves into the issue of model parameterisation, and devotes some space to investigating the measurement properties of co-occurrence statistics, namely accuracy and reliability.

The goal of Chapter 3 is to present empirical evidence for the semantic nature of the information contained in co-occurrence statistics. This is realised through two case studies which assess the psychological validity of a measure of distributional similarity. This measure, Contextual Similarity, is evaluated first against elicited semantic similarity judgements, and second through simulation of the results of a recent semantic priming experiment. Both studies have interesting implications for more conventional interpretations of the cognitive behaviours in question.

In Chapter 4, we modify the measure of representational similarity in an attempt to account for a range of context-dependent phenomena; this modification involves weighting the measure to reflect the salience of those aspects of meaning shared by the words in the context. We then show how this method can be applied to three

[4] We prefer the term *lexico-semantic*, as it is apparent that these representations contain knowledge that can be described as 'lexical' or 'semantic' (among other terms). The results of the priming simulation reported in section 3.3 lend support to the adoption of a general term.

domains: extracting the contextually appropriate meaning from an ambiguous word, characterising the effects of contextual constraint, and modelling feature priming. Chapter 4 concludes with a discussion of the inherent limitations of the representational approach to modelling context-dependent processing behaviour.

The second part of the thesis (Chapters 5-7) supplants the representational approach (where effort corresponds to distributional similarity) to modelling lexical processing behaviour, in favour of the expectation-building strategy proposed in section 1.1.1 above. Although drawing upon the same source of environmental statistics, we claim the information-oriented model to be preferable, as it allows a parsimonious account of both context-independent and context-dependent phenomena.

Chapter 5 focuses on formalising and validating the information-theoretic measure of contextual distinctiveness (CD). We hypothesise that this measure of environmental complexity has a behavioural correlate in the effort involved in processing isolated words. The rest of the chapter is largely devoted to testing this hypothesis. This is done by conducting two visual word recognition experiments, and through reanalysis of a presentation naming study carried out with Alzheimer's patients. Chapter 5 also explores the relationship between CD and other lexical properties such as frequency and ambiguity.

In Chapter 6, the emphasis is on modelling the incremental process of building semantic expectations. We motivate and implement a Bayesian update rule in order to integrate the influence of the preceding linguistic context into expectations about the meaning of upcoming words. As well as accommodating the incremental nature of comprehension, the information-oriented ICD model accounts for variability in lexical processing effort observed using the semantic priming paradigm, and predicts the effects of contextual constraint on eye movements during reading.

Chapter 7 extends the empirical assessment of the ICD model to the domains of sentence priming and multiple-priming (the influence of two or more prime words on target word processing). We apply the model to the question of the source of sentence priming – whether facilitation of a target word is due to its relationship with individual words in the sentence or with a high-level conceptual representation of the context. The model performs equally well in explaining multiple-priming effects, providing further support for its contribution towards a computational-level explanation of lexical processing effort.

In Chapter 8, we summarise the main conclusions drawn from the thesis, and outline some ideas for related future work.

2.
Representing Word Meaning

In this chapter we introduce a recent approach to the problem of modelling word meaning. Considering the meaning of words to be intimately tied to their contexts of use has a long history in the field of distributional linguistics (*eg*. Harris, 1954), but representing word meaning *quantitatively* in terms of simple co-occurrence statistics is a relatively new and fruitful direction of inquiry in psycholinguistic research. After presenting the methodology underlying the *semantic space model* approach to representing meaning, we continue with a comprehensive discussion of the issues that need to be addressed when constructing such a model. Finally, we discuss the accuracy and reliability of co-occurrence statistics, and assess the impact of some inherent properties of language corpora on vector representations created from this source of information.

2.1 Semantic space models

High-dimensional semantic space models are useful metaphors for the representation of word meaning. Word meaning varies along many dimensions; these models attempt to capture this variation in a coherent way, by locating words in a geometric space. The main principle governing their placement in the space is that words that are similar in meaning should be positioned closer together than words that are dissimilar in meaning. Although some authors have proposed the use of *features* of the type [+HUMAN] and [-CAN FLY] to define the dimensions of the semantic space (for an overview of feature-based theories of meaning, see McNamara & Miller, 1989), there has been a surge of recent research activity into the

... and this is the last thing the field anthropologist who tries to understand other peoples' symbols can afford to do. he must start with the explanations and commentaries which his informants themselves offer about their symbols. these must **first** be examined in the contexts in which they are usually employed, where they occur naturally, although subsequent generalizing discussion helps the anthropologist to improve his very superficial initial understanding. **learning** the meaning of symbols is part of the anthropologist's practical semantics: **discovering** the meaning of words, noticing when their use is appropriate and when it is not. all this requires imagination, patience, considerable linguistic skill, but above all a rigorous respect for the facts. these must come **first**; fantasy can come later ...

Co-occurrence matrix:

	these	*meaning*	*the*	*practical*	*come*
first	2				2
learning		1	1		
discovering		1	1	1	

Figure 2-1. Creation of five-dimensional co-occurrence vectors for the target words **first**, **learning** and **developing**, using a context window size of ±3 words. The text fragment is from the British National Corpus.

construction of models where the positions along each axis corresponds to *lexical co-occurrence frequencies* extracted from a large corpus of natural language. In this type of model, a word is represented as a vector, where the components of the vector are labelled with other words (the *context words*), and the value of each vector component encodes the number of times the word of interest co-occurs with the component label, within a pre-defined *window* of words. The co-occurrence vector for a word can be considered to be a high-dimensional summary of its contextual behaviour.

More formally, each word *w* in the lexicon is represented by a *k* element vector reflecting the local distributional context of *w* relative to *k* context words $c_{1...k}$ and window size *wsize*. The value of each vector element is a function of the number of times each *c* occurs within the *wsize* words before and after every instance of *w* in a large corpus. Since *k* context words define a *k*-dimensional space, *w* can be viewed as a point in semantic space, and a distance measure applied to any two words reflects their distributional, and arguably, their semantic similarity.

To illustrate the model building process more clearly, consider the fragment of text from the British National Corpus displayed in Figure 2-1. Occurrences of the

words of interest (the *target words*) **first**, **learning**, and **discovering** in the corpus fragment are indicated in boldface. In order to create vectors for these target words (setting the size of the context window to three words before and three words after the target, and considering {*these*, *meaning*, *the*, *practical*, *come*} as the set of context words), we need to examine the words appearing in the window around each of their occurrences (indicated by a background screen).

In this corpus fragment, two valid context words, *the* and *meaning*, appear in the window around both **learning** and **discovering**; *practical* additionally co-occurs with **discovering**. We therefore increment the cells of a co-occurrence matrix that correspond to these target-context word combinations. There are two occurrences of the target word **first** in the fragment; *these* co-occurs with both instances, and *come* co-occurs twice with the second instance; note that both the (**first**, *these*) and the (**first**, *come*) cells of the matrix are incremented twice. After carrying out this procedure over millions of words, the resulting vectors of co-occurrence counts for these three target words might resemble those in (1) below.

(1)

first	<113, 6, 139, 32, 66>
learning	<42, 20, 255, 16, 13>
discovering	<38, 19, 265, 2, 9>

From this contrived example we can see the appeal of the approach: the co-occurrence patterns for **learning** and **discovering** are more similar to each other than to the vector for **first**, which corresponds to intuitions about the meaning relationships between these three words.

Two intrinsic properties of semantic space models make them attractive for modelling psycholinguistic behaviour. The first is *objectivity*; semantic space models are derived from the simple distributional information contained in a record of natural language output. Minimal assumptions about the data are made (namely the identification of word boundaries), and typically no linguistic knowledge is used in their construction. It is scientifically preferable to make any model building procedure as objective as possible, and exploiting the statistical structure in the linguistic environment for a model of semantic representation eliminates the inevitable variability in, for example, the postulation and selection of features that are necessary components of most non-objective theories of word meaning.

The second property is *language-independence*. The procedure described above for tracking and recording lexical co-occurrence frequencies can in principle be done for

any language, regardless of orthographic conventions. Similar sorts of distributional analyses have been carried out using corpora of French, German and Mandarin (Redington, Chater, Huang, Chang, Finch & Chen, 1995), and results are comparable to those found with English (*eg.* Finch & Chater, 1992). The general principle that a record of language output can be exploited to create psychologically interesting lexical representations seems to hold cross-linguistically.

2.2 Corpus choice

Construction of a semantic space model requires a large amount of natural language output; the choice of this corpus certainly has an impact on the psychological plausibility of the resulting model. The corpus used for the majority of the research reported in this thesis is the 10.3 million word spoken language part of the British National Corpus (BNC; Burnage & Dunlop, 1992). The spoken subcorpus (henceforth *BNC-spoken*) consists of a mixture of speech genres sampled from demographic and context-governed sources. The largest corpus of transcribed speech previously available for research use is the London-Lund corpus (LLC; Svartvik & Quirk, 1980), which, at approximately 200,000 words, is a few orders of magnitude smaller than the BNC-spoken. Unlike the LLC, the BNC-spoken is large enough to yield reliable measurements of inter-word relationships for a substantial part of the lexicon.

2.2.1 Spoken language corpora

Spoken language corpora are characterised by substantially fewer word *types* than comparable amounts of written text (Dahl, 1979), and the higher frequency types in spoken language account for more of the *tokens*. Figure 2-2 illustrates this difference in the type:token ratio between the BNC-spoken and a comparably-sized subcorpus of the written language part of the BNC. The hyperbolic relationship between word frequency and frequency rank illustrates the most well-known of the laws described by Zipf (1935); the steeper curve for the spoken language corpus reflects its smaller type:token ratio.

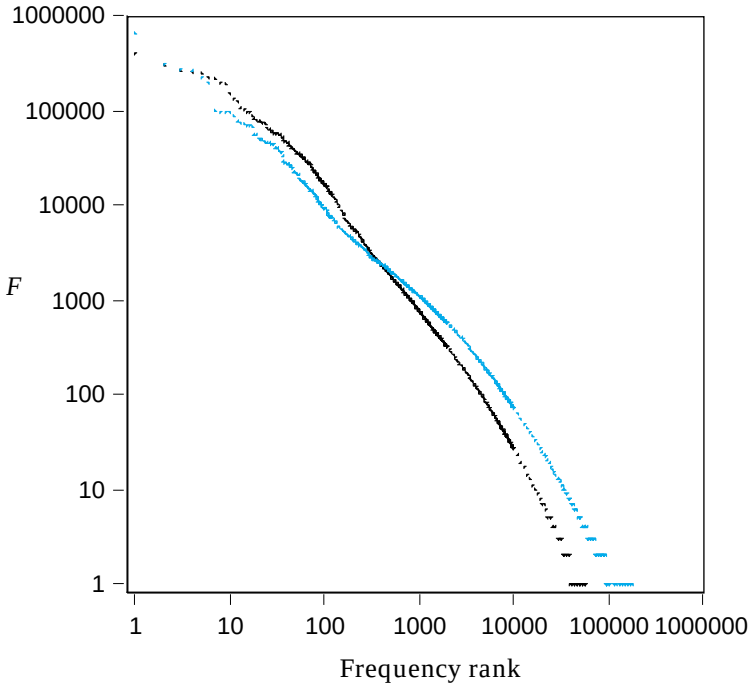

Figure 2-2. Log plot of lexeme frequency against frequency rank. The black points correspond to the BNC-spoken, and the grey points to a comparably-sized written text subcorpus of the BNC.

We chose to use spoken language as the preferred source of distributional statistics for three reasons. First, spoken language forms the primary environment for human language learning. Children's exposure to speech compared with written text is crucially much larger, and is exclusively so before they learn to read. Although vocabulary size undoubtedly increases more rapidly through reading, the core vocabulary items and their contexts of use are acquired through the vast amount of speech that children hear. Words have much more opportunity to be learned through spoken language than through written sources, and their semantic representations would be expected to reflect spoken language context to a greater extent. Second, because of the smaller type:token ratio for spoken language, a single word type is encountered, on average, an order of magnitude more often than a single word type in written language. This means that spoken language, in general, provides a more reliable source of contextual information for a given word, which is advantageous for both learning word meaning and for constructing reliable co-occurrence representations.

The third motivation for preferring spoken language over written is empirical; we shall see in Chapter 5 that a semantic space model constructed using the BNC-

spoken subcorpus provides a closer fit to behavioural data than if based on a comparably-sized corpus of written text. Of course, strong empirical evidence for the superiority of the spoken language source would provide the most compelling justification, but this distinction is only of peripheral concern to this thesis and will not be explored in depth. We shall employ both spoken and written sources of language output in the simulations reported in the following chapters.

2.2.2 Corpus preparation

The BNC-spoken subcorpus consists of 863 'texts', which were transcribed from diverse sources such as unscripted informal conversation, radio programmes and government meetings, in order to represent a wide range of contemporary spoken British English. To prepare the corpus for the computational analyses carried out in this thesis, we took the following steps.

First, the corpus was filtered to remove punctuation and SGML markup, retaining only the words together with their part of speech tags. Next, uppercase and lowercase type were conflated to lowercase. Finally, the corpus was lemmatised (see also section 2.3.1) by mapping each word form to its corresponding *lexeme* in the CELEX lexical database (Baayen, Piepenbrock & Van Rijn, 1993), and then replacing the word in the corpus with its lexeme's canonical form.[1] This was done by mapping the tagset used to assign a part-of-speech label to each word in the BNC, to the much smaller set of part-of-speech categories employed by CELEX. This procedure meant that ambiguous forms such as *broke* (which can occur as an adjective or as the past tense form of the verb *break*) were disambiguated for syntactic category. For example, *broke* was replaced with either *broke* or *break*, depending on its part-of-speech tag. In the case that the CELEX database did not list the corpus item, it was retained as is. Since the BNC tagger assigned a large number of 'ambiguity tags' (4.7% of the BNC) when it could not decide between two simple tags (*eg*. AJ0-VVG for gerunds such as *walking*), the lemmatiser also could not decide, and by default the first match found in CELEX was retrieved.

[1] The canonical (or citation) form is the form typically used as the headword of a dictionary entry: in English, the singular for nouns, the first person singular present tense for verbs, and the positive form for adjectives. Note that our use of *lexeme* differs from that of Cruse (1986), who considers two identical word forms with no predictable meaning relationship to belong to two different lexemes. And unlike CELEX's definition of *lemma*, our use of lexeme collapses together different syntactic categories sharing the same canonical form, such as the noun *walk* with the verb *walk* in all its inflected forms.

After corpus preparation, the 1,032,581 'sentences' in the BNC-spoken consisted of 10,286,448 lexeme tokens, representing 45,451 unique lexeme types.[2] Note that the BNC tokeniser treats multi-word units such as *at least*, *a bit* and *out of* as single words (marked up with <W></W> delimiters); this segmentation was retained.

2.3 Model parameters

There is a large parameter space to consider when building a semantic space model; the size and shape of the context window, the number and selection of context words, and the size of the corpus all affect the quality of the vector representations extracted. Since one normally wants to create an 'optimal' representational model, how can the 'best' parameter settings be determined? Exploration of the parameter space is typically done in computational linguistics research in terms of end-task performance. That is, the problem that the model was designed to address is also used to set the parameters. Parameter settings are adjusted until performance on the end-task has peaked or reached some satisfactory level. This means that model parameters end up being optimised for the task at hand, *ie*. the model overfits the data. Landauer and Dumais (1997) even attribute explanatory power to this approach for their semantic space model of vocabulary learning, arguing that there is no *a priori* reason why a parameter setting should give optimal modelling performance, unless that particular setting reflected the cognitive mechanism involved.

Using performance on a single end-task to set parameters seems deficient for purposes of psychological modelling, where generality across a range of language processing behaviours is desirable. For example, a semantic space model optimised to capture word-to-word similarity might be seriously deficient when applied to another phenomenon. Evaluation of the effect of parameter changes should clearly be done using more than one task. Levy and colleagues (Levy, Bullinaria & Patel, 1997; Patel, Bullinaria & Levy, 1998) have recently reported systematic searches along several dimensions of the parameter space, and conclude that parameter settings have a substantial effect on the quality of the vector representations, as

[2] The BNC web pages (`http://info.ox.ac.uk:80/bnc/what/balance.html`) quote 1,042,397 S-units (text marked up with <S></S> delimiters), and 10,365,464 <W></W> delimited items. The discrepancies between these figures and ours are likely due to errors introduced while stripping off the SGML markup.

measured by several evaluation benchmarks, such as semantic category norms and synonym choice tests.

We propose that independent criteria should be used as often as possible to set model parameters. This means using the results of relevant psycholinguistic research to motivate and support the parameter settings chosen. Appendix A reports the results of varying several model parameters, using elicited semantic similarity ratings (see Chapter 3) as the evaluation standard.

2.3.1 Lemmatisation

All of the corpus-derived measures used in this thesis – word frequency estimates and measures involving high-dimensional vector representations – draw upon corpus counts for *lexemes*, rather than *surface word forms*. As a result of lemmatising the corpus, the counts for all inflectional variants of a word are collapsed together into a single lexeme count. For example, *walk*, *walking*, *walked* and *walks* all share their high-dimensional vector representation, labelled with their canonical form *walk*; and similarly there is only one dimension label <walk> corresponding to all four variants. Other types of inflectional morphology conflated by lemmatisation are noun plural suffixes (*eg. cats*), and comparative and superlative adjective forms (*eg. cleaner, cleanest*). Lemmatisation was motivated by the observation that meaning is normally preserved across the inflectional variants of a lexeme, whereas derivational morphological variants are often semantically opaque.

There is some evidence that human lexical processing draws upon lexeme frequency (also referred to as *stem* or *summed word form* frequency) information, in preference to surface word form statistics. Early word recognition studies (*eg.* Bradley, 1983; Taft, 1979) demonstrated lexeme frequency to be a better predictor of processing time than simple surface frequency. For example, although *shoe* and *fork* are matched for corpus frequency, recognition of *shoe* is faster than *fork* because *shoes* is much more frequent than *forks* (Taft, 1979). This finding suggests that the basic unit of lexical representation is the lexeme, rather the surface word form.

More recently, Baayen, Dijkstra and Schreuder (1997) showed that lexical decision latencies for singular Dutch nouns of differing surface frequency were statistically equivalent when the materials were matched for lexeme frequency. However, this was not the case for plural nouns, for which surface frequency effects were found. Baayen *et al.* propose that it is more efficient for certain morphologically complex words to be stored 'whole' at some level of representation,

due to orthographic form ambiguity. For instance, many noun plurals (ending in *-en*) are ambiguous with verbs that share the stem and also use *-en* to mark the infinitive or past participle. Although Baayen *et al.*'s work indicates that for certain categories the basic unit of representation may actually be the surface form, our method simplifies the issue by lemmatising all morphologically complex words.

All of the psychologically-oriented research using semantic space models that we are aware of (*eg*. Huckle, 1996; Landauer & Dumais, 1997; Lund & Burgess, 1996; Redington, Chater & Finch, 1998) treats the surface word form as the basic representational entity. In support, Huckle (1996) argues that making minimal assumptions about the data is of primary importance. However, treating the surface word form as primary implies the presence of unnecessary redundancy in the semantic lexicon, which is not consistent with the psycholinguistic evidence reviewed here.

2.3.2 Window size

The size of the context window used to define co-occurrence has a substantial influence on a semantic space model's fit to psychological data (Levy *et al*., 1997). Research based on information retrieval techniques typically employs extremely large windows, sometimes even the entire text (*eg*. Landauer & Dumais, 1997; Schütze, 1992, 1993, 1998); conversely, work investigating the role of distributional information for syntactic category acquisition tends to use small windows (*eg*. Redington, Chater & Finch, 1998). Levy *et al*. found that small window sizes (around two words to either side of the target) worked best for their synonym choice evaluation task, but slightly larger windows were better (around ±4 words) for their semantic categorisation task.

The context window can be considered to correspond to aspects of the linguistic environment to which people attend when processing a word. The limitations of short term memory thus conceivably constrain the number of words used for the size of the 'previous' context window. Huckle (1996) provides a rough calculation of the number of words that, on average, would fit into the short term store (or *phonological loop* [Baddeley, 1990]); he suggests that 8.8 words is a reasonable estimate. This provides support for the success of those models outlined above that employ small windows. The window size parameter for all of the semantic space models presented in this thesis is therefore set between three and five words before and after the target word.

The use of 'forward' context for creating vector representations receives some support from a study of spoken word recognition in context using the gating paradigm (Bard, Shillcock & Altmann, 1988), which demonstrated that approximately 20 percent of words could not be identified until subsequent context had been presented. Using forward context also has purely empirical justification in terms of model fit (see Appendix A).

We chose to ignore linguistic boundaries such as sentence beginnings and speaker changeovers when moving the context window over the corpus. Preliminary experiments indicated that taking such boundaries into consideration has little effect on the psychological plausibility of the resulting vector representations.

2.3.3 The function-content word distinction

We made a psychologically-motivated decision regarding the choice of context words for semantic space model construction; we exclude the class of function words from consideration as vector components. There is a growing literature[3] on the processing and representational differences between functional and contentive expressions; for example, they are distinguished in formal linguistic theory (Cann, *in press*), by lexical priming behaviour (*eg*. Shillcock & Bard, 1993), EEG patterns (Pulvermeyer, Lutzenberger & Birbaumer, 1995), frequency sensitivity (Bradley, 1983; but see Segui, Frauenfelder, Lainé & Mehler, 1987), and phonetic realisation (*eg*. Cutler, 1993). In addition, the function word class seems to be selectively impaired in certain types of aphasia (Swinney, Zurif & Cutler, 1980), which is suggestive of major representational differences between the two classes, and makes distinguishing them in a model of lexico-semantic representation psychologically attractive. Finally, there is empirical justification for excluding function words when defining the dimensions of a semantic space model; semantic similarity relations are simply modelled more closely (see Appendix A).

From the point of view of achieving plausible *semantic* representations for words, excluding function words intuitively reduces the influence of semantically arbitrary factors, namely those dimensions of variation that are chiefly concerned with grammatical properties. For example, the fact that the determiner *the* co-occurs with *chair* does not distinguish the meaning of *chair* from *mouth*, since *the* can co-occur equally well with *mouth*; what this information can do is make *syntactic* distinctions,

[3] See Cann (*in press*) for a comprehensive review of the differences between the two classes.

such as distinguishing nouns from verbs. Function words can be viewed as the 'building blocks' of syntax, and indeed the use of a context word set consisting primarily of function words to define the axes of the high-dimensional space has led to successful induction of syntactic categories from distributional statistics (*eg*. Redington, Chater & Finch, 1998). Of course, semantically relevant roles can often be attributed to certain function words in minimal contexts – the different prepositions in the sentences "He ran *through* the door." and "He ran *into* the door." certainly highlight different aspects of the meaning of *door*. But since either preposition can occur in the context window around *door*, a record of co-occurrences with *through* and *into* does not necessarily contribute to distinguishing the meaning of *door* from other words.

In order to define the set of function words, it is desirable to limit subjective influence, since opinions differ on the functional-contentive divide, particularly when approaching the boundary between the two classes. For example, conjunctions and articles might be nearly indisputable examples of function word categories, but what about indefinite pronouns such as *anyone*? Prepositions, though having a salient syntactic function, also seem to carry some semantic weight, such as directionality (*eg*. *through*, *out of*).

An additional classification problem becomes apparent due to lexical ambiguity; a particular orthographic form might map to more than one syntactic category. For example, *down* is listed in the CELEX lexical database as an adverb, preposition, adjective, verb and noun; should *down* be classified as a function or content word? A compromise solution was sought for the purposes of this thesis. First, we considered four of the categories used by CELEX to represent the class of function words: articles (CELEX Class #5), pronouns (#6), conjunctions (#9), and prepositions (#8). Next, membership of this set was reduced by filtering out the ambiguous forms (such as *down* and *off*) that were more frequently used as content words, according to CELEX. To the resulting set we appended the set of auxiliary verbs such as *may* and *have*, and the copula *be*. Finally, a small number of function words that occurred in the corpus, but not in CELEX, were added to the list. All of these were obvious conjunctions, such as *as soon as* and *so that*. The final set of function words defined using this procedure appears in Appendix B.

2.3.4 Choice of context words

The *number* and *selection* of the context words used to define the dimensions of the space are also important parameters. Previous semantic space research has typically chosen the set of context words by simply selecting the topmost *k* items from a frequency list created from the corpus. For reasons of reliability and storage, *k* is typically fairly small, around 100-2000 (*eg*. Huckle, 1996; Redington, Chater & Finch, 1998), or may start off large, perhaps with *k* initially set to the number of corpus types, but be subsequently dropped through dimensionality reduction techniques (*eg*. Schütze, 1998).

Levy *et al*. (1997) provide plots which indicate that performance on both their categorisation and synonym choice evaluation tasks rises as the number of context words increases (holding the window size constant), when using the simple frequency list approach to vary the number of context words. This method was used for most of the computational modelling reported in this thesis, but with a frequency list restricted to content words. The number of context words, and thus the dimensionality of the representational space, was determined empirically (*cf*. section 5.1.2).

2.3.5 Encoding co-occurrence

The usual approach to representing co-occurrence information in a word vector is to simply encode the corpus co-occurrence frequency itself. However, depending on the eventual application of the semantic space model, this may not be the best method. For example, if 'semantic distance' is to be estimated between pairs of words using Euclidean distance, the metric will be inaccurate for words that differ in corpus frequency, since the co-occurrence counts involving high-frequency target words tend to be larger than the counts involving low-frequency targets. One way around this problem is to *normalise* the vectors; for instance by encoding the *conditional probability* of observing the context word in the window, given the appearance of the target, or by normalising the Euclidean length of the word vectors to a constant. Normalisation allows meaningful comparisons to be made between different distance or similarity measures.

In the semantic space simulations reported in Chapters 3 and 4, we employed potentially more informative values than simple co-occurrence counts as vector components. The motivation is as follows: if we are interested in comparing the

vector representations for two words, and the co-occurrence counts for both words with a particular context word are identical, then this dimension does not contribute towards distinguishing the two words when perhaps it should. For example, in a hypothetical corpus of ten million words, *quite* and *extremely* each might co-occur 50 times (in a context window of one word before) with the context word *interesting*, which has, say, a corpus frequency of 1,500. However, because *quite* has a corpus frequency of 11,000, and *extremely* has a much smaller frequency of 500, the fact that *extremely* and *interesting* co-occur as often as they do is intuitively surprising.

The log-likelihood statistic seems appropriate for capturing this intuition. It can be considered as an estimate of how 'surprising' a particular co-occurrence count is, given the additional knowledge of the independent counts of the members of the pair. The log-likelihood ratio (a goodness-of-fit statistic) was introduced into corpus analysis in a frequently cited paper by Dunning (1993) as a measure of the association between two co-occurring words. In brief, this statistic tests the independence of the counts for each member of the co-occurrence pair. Dunning points out that the log-likelihood ratio allows a meaningful comparison to be made between rare and common events, unlike other proposed measures of word association, such as pointwise mutual information (Church & Hanks, 1990).

For the example presented above, the log-likelihood ratios for the target-context word pairs <*quite, interesting*> and <*extremely*, *interesting*> are 246.23 and 557.24, respectively, which confirms the intuition that the latter co-occurrence frequency is the more surprising, even though the counts are identical.[4] The effect of using this statistic instead of simple counts in a semantic space model is to 'push apart' the representations for words which might be very close along certain dimensions. The log-likelihood ratio offers an alternative way to quantify the lexical association between words; unfortunately, the only reasonable way to justify its use psychologically appears to be through empirical means (see Appendix A).

2.3.6 Distance/similarity measures

A central motivation for creating high-dimensional lexical representations is to be able to quantify the relationships between them, either directly, using a 'semantic

[4] Note that the conditional probabilities $P(interesting|quite)=0.00455$ and $P(interesting|extremely)=0.1$ also encode this difference. However, the latter event is 22 times more likely to occur than the former, as opposed to being 2.26 times as 'surprising' if their log-likelihood ratios are compared.

distance' measure, or indirectly, using descriptive statistical procedures such as hierarchical cluster analysis. There are many ways to compare two vectors, but the most common methods used in semantic space research are geometric measures of distance and similarity, such as Euclidean distance, City-block distance, and the cosine of the angle between vectors. Other commonly-used semantic distance measures are relative entropy (or Kullback-Leibler divergence), which is an information-theoretic measure of the dissimilarity of two probability distributions, and the Spearman rank correlation coefficient.

While technically not a model parameter, the choice of vector comparison method certainly has impact on how the relationships between words are interpreted, and the choice of measure is not independent from decisions made about other model parameters. The choice of semantic distance measure is perhaps best made empirically, as is routinely done in the field of computational linguistics (*eg*. Dagan, Lee & Pereira, 1999). In the work reported in Chapters 3 and 4, we adopt the cosine of the angle between word vectors as an estimate of semantic distance:

$$cosine(x,y) = \frac{\sum_{j=1}^{n} x_j y_j}{\sqrt{\sum_{j=1}^{n} x_j^2}\sqrt{\sum_{j=1}^{n} y_j^2}}$$

As applied to co-occurrence vectors, the cosine of the angle between the vector for word *x* and the vector for word *y* has a minimum value of zero, if the word vectors are orthogonal, and a maximum value of one, if they point in exactly the same direction. Note that although the semantic distance between vectors for two words of different frequency can be measured without normalising the vectors (since only the direction of the vectors is compared), the cosine is nevertheless sensitive to vector *sparseness*. Co-occurrence vectors for low-frequency words tend to be more sparse (*ie.* contain fewer non-zero components) than vectors for high-frequency words, and the cosine tends to decrease as vector sparseness increases. Consequently, the cosine measure is best used when comparing the vectors for words of similar frequency.

2.4 Reliability and accuracy of vector representations

Before proceeding with an inquiry into the psychological validity of co-occurrence information (Chapter 3), a few words should be said about the quality of the vectors created using the current methodology. There are two distinct, but related issues here. The first is *reliability*. Reliability addresses the question of replicability using a different source of data: assuming all else to be equal, would one obtain identical or nearly identical vector representations for a particular word from two different corpora? The second is *accuracy*. How representative is a particular co-occurrence vector of the 'true' vector that would be obtained given access to an infinite corpus? One major factor affecting accuracy is known as the problem of *data sparseness*, to which we turn next.

2.4.1 The sparse data problem

It is clear that the procedure of collecting co-occurrence counts from a corpus and then constructing word vectors from these counts suffers from the standard problems of point estimation. A co-occurrence count is a statistic from a finite *sample* of language (the corpus) used to estimate a population *parameter*. The co-occurrence *probability* of a context word c_i given observation of a target word t can correspondingly be estimated using the sample *relative frequency*, adjusted for the size of the context window (*wsize*):

$$P(c_i|t) = \frac{f(c_i,t)}{f(t)} \cdot \frac{1}{wsize}$$

This is also known as the maximum likelihood estimator (MLE), because the probability of the observed data is maximised. Because the sample relative frequency is an estimate, it suffers from the usual ailments of parameter estimation due to sampling issues. Sample relative frequencies are *biased* estimators of population probabilities, since all of the probability mass is distributed amongst the events which have actually been observed in the sample. As a result, low relative frequencies tend to be inflated estimates of their corresponding population values (Carroll, 1970). More critically, the MLE for an *unobserved* event is zero – likely lower than the event's 'true' probability. This is the core of the sparse data problem; a relative frequency of zero in the available corpus does not mean that the word

pair would *never* occur together given more data, and hence the MLE is not an accurate estimator of co-occurrence probabilities.

One obvious solution to improving the MLE for co-occurrence probabilities involving low-frequency words is to increase the sample size (*ie.* make the corpus bigger), which should have the effect of reducing the amount of noise in the counts. However, although the vectors for words considered to be rare in the original corpus might now be more accurate, the tail of the frequency distribution is of course still present in the larger corpus.

2.4.2 The 'burstiness' of words

Note that data sparseness does not necessarily affect reliability; co-occurrence vectors for rare words could, in principle, be quite similar when constructed from different corpora. But one inherent property of corpora with the potential to affect both reliability and accuracy is related to the following observation. A corpus is made up of texts, and each text is typically about a topic. The topic tends to be characterised by a small set of low frequency words, and consequently these words recur within the same text more often than expected by chance. This phenomenon has been termed the *burstiness* of words (Katz, 1996), or *word density* (Dennis, 1994). Because of this natural property of corpora, a co-occurrence count might reflect the characteristics and peculiarities of a particular text rather than its 'true' count in the population of natural language, which would have direct impact on a word's vector representation.

How can this problem be avoided (assuming that it is, indeed, a problem)? One possibility is to take the *dispersion* of the event into consideration (*eg.* Carroll, 1970). The logic is as follows: if the co-occurrence pair is dispersed more or less evenly throughout the texts that make up the corpus, its relative frequency is likely to be an accurate estimate of the population value. If the co-occurrence pair is not scattered evenly, then a dispersion measure will reflect this, and the relative frequency can be re-estimated by adjusting the co-occurrence frequency downwards by an amount inversely proportional to its dispersion score. Low frequency events typically receive a low dispersion score, which often results in an adjusted frequency of zero; it appears that this procedure undesirably increases vector sparseness.

2.4.3 Estimating reliability using external evidence

Before legitimatising the use of co-occurrence vector representations in psychological modelling, it is necessary to ensure that the vectors are *reliable*. The most straightforward approach for determining the reliability of a particular co-occurrence vector is to construct vectors for the same word from two or more corpora, and compare the vectors. Broadly speaking, vector representations for the same word derived from the distributional information contained in different corpora should point in the same direction. However, this approach to assessing vector reliability was not feasible, since no speech corpora of comparable size to the BNC-spoken were available at the time of writing. One can, however, split a single corpus in half and compare the vectors created from each half to each other. Using this method, we estimated the reliability of the vector representations for words selected randomly from a range of frequency strata.

The BNC-spoken was divided into two halves by alternating 10,000 token chunks: the first half consisted of 5,143,107 words, the second half contained 5,143,341 words. Natural log-transformed lexeme frequency in the 10M word BNC-spoken ranged from 0 to 12.963 (see Table 2-1). We divided this range into 8 equally-spaced intervals, and generated a random sample of 100 words from each bin, with the constraint that the selected words had to appear in both subcorpora. (Bins 1 and 2 consisted of 21 and 69 words, respectively.) Next, we extracted co-occurrence vectors for each word (690 in total) from both subcorpora, using a window size of ±5, and the 500 most frequent content words in the BNC-spoken as context words.

Table 2-1. Mean Reliability of Co-occurrence Vectors for Eight Samples.

Bin	Log Frequency Range	1st Word in Bin	N	Sample Size	Mean *W*	X^2
1	12.963-11.344	be	21	21	0.988	986.25*
2	11.343-9.724	yeah	69	69	0.965	963.48*
3	9.723-8.103	work	195	100	0.902	899.76*
4	8.102-6.483	case	810	100	0.814	812.06*
5	6.482-4.862	goal	2204	100	0.708	706.34*
6	4.861-3.242	valid	4624	100	0.619	617.91*
7	3.241-1.621	zebra	9150	100	0.563	562.01
8	1.620-0.000	zulu	28378	100	0.557	555.66

* Significant at α=0.01

In order to estimate the reliability of the vector representations in each bin, we used Kendall's coefficient of concordance *W* to measure the degree of agreement between the two subcorpora. The data points were the co-occurrence counts for the word of interest with each of the 500 context words. Kendall's *W* will be high when the component values of the vectors for the same word created from more than one corpus have similar rank orders. Mean *W* scores for the words in each bin are given in Table 2-1.

The X^2 statistic was used to test the significance of mean *W* for each bin. From Table 2-1 it is clear that the reliability of a vector representation decreases with corpus frequency; at the α=0.01 level of significance, mean *W* for the two lowest frequency bins could not be confidently distinguished from chance. Therefore, it appears that a useful lower bound of corpus frequency can be determined for the construction of co-occurrence vectors (at least under the current parameter settings). Since the minimum log frequency for bin 6, 3.242, corresponds to a frequency of 25, we decided to not use words with a lexeme frequency of less than this value in any of the computational modelling reported in this thesis. Although restricting the usable vocabulary in the BNC-spoken to approximately 8,000 lexemes, applying this frequency threshold reduces concerns about vector reliability.

2.4.4 Could smoothing improve accuracy?

In order to ensure that co-occurrence vector representations are *accurate*, we would like to obtain the best estimates possible of the 'true' co-occurrence probabilities; this is especially important for target-context word pairs which do *not* co-occur in the corpus. *Smoothing* is a technique widely used in statistical language modelling to address the sparse data problem (*eg*. Church & Gale, 1991; Dagan, Lee & Pereira, 1999); it is done primarily to provide data about unobserved events (where the MLE is clearly inaccurate). The question to be addressed in this section is "could smoothing methods improve the quality of vector representations?"

In statistical NLP, the usefulness of smoothing depends on the application. A number of approaches have been proposed in the literature, all sharing the basic goal of adjusting the sample relative frequency in order to more accurately estimate the 'true' probability. A common approach to smoothing is to interpolate a bigram relative frequency with a unigram relative frequency, under the assumption that the zero count observed for the bigram is due to the omission of the bigram in the sample, and not to its non-occurrence in the population.

It is clear that a successful smoothing procedure needs to distinguish between unobserved events that are likely to occur in language from those that are not. This requires information over and above the distributional information contained in a corpus, and existing smoothing methods typically do not use any additional (*eg*. linguistic) knowledge. Consequently, a trade-off with these methods becomes apparent: is postulating the existence of an event, when in fact it does *not* occur in the population, a more or less serious error to make than retaining a zero value, when it is merely the result of insufficient evidence? For example, one probably never finds a contiguous co-occurrence of two adjectives representing opposite ends of a property scale, such as *freezing* and *hot*. Smoothing this zero probability, $P(freezing|hot)=0$, by interpolation with the relative frequency of either *freezing* or *hot* will clearly overestimate the population value, due to ignorance of the semantic constraints on word combination. An effective smoothing method should not compromise linguistic plausibility when 'recreating' co-occurrence counts.

Smoothing of the sort that recreates co-occurrences with a 'syntactic' function, such as postulating a nonzero co-occurrence probability for *the* with *troglodyte*, when the sample relative frequency $P(the|troglodyte)$ is zero, would seem to be orthogonal to the problem raised above. However, because the representational models investigated in this thesis do not use function words to label the dimensions of the space, all co-occurrences necessarily involve content words, and so conventional smoothing procedures would be prone to create the sort of plausibility error described above.

The recent approach to smoothing taken by Dagan and colleagues (Dagan, Lee & Pereira, 1999; Dagan, Pereira & Lee, 1994) has the potential to avoid this problem. Their approach, called *similarity-based estimation* of co-occurrence probabilities, is specifically aimed at distinguishing between word combinations that are likely to occur in language from those that are not. For example, in order to estimate the conditional probability of *peach* given the verb *eat*, $P(peach|eat)$, when the relative frequency of the pair is zero in the corpus, they exploit relative frequency data for co-occurrences of *peach* with the set of words Y that are distributionally similar to *eat*. One measure of distributional similarity employed by Dagan *et al*. is a weighted function of the Kullback-Leibler distance between the conditional probability distributions $P(X|eat)$ and $P(X|Y)$. In brief, Y is distributionally similar to *eat* if they both tend to co-occur to the same extent with other words. Presumably this method would identify verbs that take the same sorts of objects as *eat*, such as *buy*, *cook*, or

serve. The co-occurrence probability *P*(*peach*|*eat*) is next estimated using information about the observed co-occurrences of *peach* with the set of words *Y* that are distributionally similar to *eat*.

Dagan *et al*. (1994) demonstrated that their similarity-based smoothing method was superior to frequency-based estimation (*ie.* smoothing based on unigram probabilities) using perplexity, a standard evaluation metric for how well a statistical language model captures a test corpus. However, it seems that this method would still fail on the *freezing-hot* example. Because antonyms tend to be distributionally similar, *P*(*freezing*|*hot*) would undoubtedly be smoothed with co-occurrences of *freezing* with *cold*, resulting in a estimated conditional probability that is greater than zero. Although awaiting empirical verification, it appears that similarity-based smoothing would be deficient for cases such as these, by effectively postulating the existence of implausible lexical relationships.

In summary, it appears that although there are certain clear disadvantages to *not* smoothing (relative frequencies are biased estimators), the advantages of smoothing methods have not yet been satisfactorily established. Because of uncertainty about their efficacy, the potential for problems such as those raised above, and also because state-of-the-art smoothing techniques are not trivial to implement, we chose not to smooth.

2.4.5 Solving the sparse data problem

The issue of vector reliability is straightforward to address: if the evidence is not reliable, do not use it. This pragmatic solution is also germane to the accuracy issue; to 'solve' the sparse data problem we simply refrain from creating vectors for low frequency words. We believe this approach is actually preferable to smoothing for purposes of modelling human language behaviour. If one could record the actual number of instances where a group of subjects have encountered or produced a very rare word, such as *amorphous* – which occurs once in the BNC-spoken – the resulting distribution would have a much higher variance than if the same procedure was carried out for a substantially more frequent word. Individual differences in experience with very rare words thus form an additional source of variation that will affect experimental measures of processing difficulty such as visual lexical decision response latency. Therefore, we should not anticipate the snapshot of contextual behaviour captured by the co-occurrence vector for a very rare word to correspond to some abstract semantic representation shared by the subject

population, because there is likely no such representation which could be described as consistent across subjects. Consequently, we should not expect accurate behavioural prediction for very rare words from a model created from simple distributional statistics.

2.5 Summary

In this chapter, we laid the methodological foundations for the computational modelling carried out in the remainder of the thesis. We presented the basic methodology behind the construction of a semantic space model from a large corpus of natural language. Next, issues of model parameterisation were raised, and we attempted to provide psychologically-motivated justifications for the selection of corpus, window size, context word set, similarity measure and other important parameter settings. We empirically established a criterion for the reliability of vector representations, which allowed a lower bound to be specified in terms of word frequency. We also described the infamous *sparse data problem* and its consequences for corpus-based modelling, and considered the impact that statistical smoothing methods might be expected to have on the accuracy of co-occurrence statistics.

3.
Psychological Validity

In this chapter we investigate the psychological reality of the information contained in a semantic space model. By assessing the *validity* of a measure applied to the relationships between co-occurrence vectors, it is possible to confirm the semantic properties of the high-dimensional representations themselves. We first discuss previous research that uses distributional statistics as a type of representational model, and then evaluate the proposed correspondence between semantic similarity and vector similarity in Experiments 1 and 2. Further evidence for the psychological validity of co-occurrence information is provided by an extensive semantic space simulation of a recent lexical priming experiment (Experiments 3 and 4).

3.1 Previous research

High-dimensional semantic space models have been of interest to researchers in the traditionally disparate fields of computational linguistics and cognitive psychology. The central motivation for their use (from both perspectives) is that the contexts that a word occurs in contain useful information about its meaning. The majority of studies have addressed the problem of *representing* word meaning, either for improving the machine processing of language or for modelling human language behaviour. Semantic space research has been driven by the assumption that a word's semantic properties are latent in its distributional pattern of use, as recorded in a large corpus. Note that these corpus-based techniques have been used to construct *representational* models – vector representations are *static* – and the

quantity of interest is the relationship (some measure of vector similarity) *between* lexical representations.

In the following, we briefly review semantic space model research addressing the diverse aims of language engineering and cognitive modelling. Common to both perspectives is the underlying principle that the relationship between word vectors (often referred to as 'semantic distance') corresponds to relatedness in meaning.

3.1.1 The computational linguistics perspective

For researchers in the natural language processing fields, statistical methods for automatically extracting semantic information from large corpora offer a potential solution to the *knowledge acquisition bottleneck*. This term refers to the time- and labour-intensive task of manually creating lexical resources for language processing applications, such as information retrieval, natural language understanding, and machine translation. For tasks deemed necessary for successful attainment of these goals, such as word sense disambiguation (WSD), a method for representing word meaning is crucial. Below, we outline three representative applications of the semantic space approach to encoding lexical semantic information.

Schütze (1992, 1993, 1998) exploits co-occurrence statistics for the task of word sense discrimination, which he describes as a subproblem of WSD; before a new occurrence of an ambiguous target word such as *train* can be disambiguated (*ie*. classified as an instance of one sense or another), the various corpus instances of *train* must first be grouped together according to sense. Following this categorisation step, the ambiguous target is assigned to the sense-group 'closest' in meaning.

Schütze's approach relies on the equation of a word's meaning with its co-occurrence vector representation. He first constructs high-dimensional word vectors (employing between 1000 and 5000 context words), and then reduces dimensionality using singular value decomposition (SVD), a mathematical technique which attempts to preserve high-dimensional inter-word relationships in a space of fewer dimensions. By averaging together the vectors for all the words in the context surrounding the ambiguous target, a word-in-context representation is created. Schütze then uses an agglomerative clustering algorithm to group these 'context' vectors together according to their similarity, providing the sense-groups into which the test items can be classified (and thus disambiguated). WSD performance was reasonable, approaching 91% accuracy using a test set of 20 ambiguous words. It is clear that Schütze's method depends on the hypothesis that words that occur in

similar contexts are similar in meaning. “By looking at the amount of overlap between two vectors, once can roughly determine how closely they are related semantically” (Schütze, 1998, p. 101).

Grefenstette (1994) investigates the use of co-occurrence information for automatic thesaurus compilation, under the working assumption that two words are related in meaning to the extent that they share attributes; he considers the relevant attributes to be the words that occur in the immediate context. Grefenstette compares two approaches to defining co-occurrence: (a) two words in a syntactic relationship, such as subject-verb or modifier-noun, and (b) a pair of words occurring within a region of text, irrespective of syntactic relationship (the conventional window-based approach). Using the groupings provided by two machine-readable thesauri as benchmarks for evaluation, he finds that a noun and its nearest neighbour (according to his semantic distance measure) were often found in the same thesaurus entry. The syntactic preprocessing step gave significantly better results than the window method, though it should be noted that the window size was quite large (defined as the 10 nouns, adjectives or verbs before and after the target word, but within the same sentence). Since the thesauri were constructed according to lexicographers’ intuitions about semantic relatedness, Grefenstette’s approach appears to be capturing some of these intuitions, supporting his working hypothesis that “… words that are used in a similar way throughout a corpus are indeed semantically similar.” (Grefenstette, 1994, p. 34).

Poesio, Schulte im Walde and Brew (1997) evaluate the utility of a semantic distance measure for the task of resolving a definite description with its antecedent. For instance, in their example “John saw a truck stopped at an intersection. *The vehicle*’s engine was smoking.”, the definite noun phrase *the vehicle* refers to *the truck* mentioned in the previous context. Poesio *et al.* test the hypothesis that the correct antecedent is closer in corpus-derived semantic distance to the noun in the definite description than to any of the other potential antecedents. Performance was much better than chance, but was generally low (<25%), which was partly due to the cases where the correct antecedent was not the word most (intuitively) similar in meaning to the head noun of the definite description.

All three projects reviewed above have capitalised on the assumption that distributional similarity (or alternatively, substitutability in context) corresponds to the psychological concept of semantic relatedness. However, only Grefenstette’s (1994) work has addressed the validity of this assumption, and then only

indirectly, by assuming that thesaurus entries are organised according to principles of (psychological) similarity.

3.1.2 The psychological perspective

From the psychological perspective, distributional statistics were initially investigated for their role in the unsupervised discovery of syntactic categories, rather than semantic relations *per se* (*eg.* Finch & Chater, 1992; Redington, Chater & Finch, 1998). However, hierarchical cluster analysis applied to the vector similarity matrix often revealed groups of words that were intuitively semantically related, providing the first indications that both syntactic *and* semantic properties of words seemed to be encoded in the high-dimensional space.

Huckle (1996) explored this direction further, in the context of modelling the acquisition of semantic categories. He used Roget's thesaurus as a benchmark in order to evaluate the 'semantic-ness' of the clustering produced using both window-based and unsupervised neural network approaches. Although the results of both methods indicated matches to Roget's category structure that were better than chance, the clustering appeared to contain as much noise as it did valid semantic categories.

Bullinaria and Huckle (1997) avoided the inherent limitations of analysing the output of a clustering algorithm by instantiating co-occurrence vector representations in a connectionist model of the lexical decision task. Their cascaded feed-forward network model successfully simulated the semantic priming effect (faster lexical decisions to a target word preceded by a semantically related prime word than by an unrelated prime), but unfortunately their simulations did not use real experimental stimuli. Instead, the authors considered a target word's three nearest neighbours (in a 200-dimensional semantic space) to be its related primes and its three furthest neighbours to be its unrelated primes. The psychological validity of Bullinaria and Huckle's semantic distance measure could only be established if priming was obtained with human subjects using the same stimuli.

Landauer and colleagues (*eg.* Landauer & Dumais, 1997; Landauer, Foltz & Laham, 1998) borrowed the term-by-document vector-space model extensively used in the field of Information Retrieval and applied it to the problem of representing word and sentence meaning. The difference between their approach (called Latent Semantic Analysis, or LSA) and the class of semantic space models where co-occurrence is defined over a context window is that in LSA, vector elements initially

encode the number of times the target word occurs in a particular *document*. Two words are thus considered similar if they tend to appear in the same documents with approximately equal frequency. Next, SVD is used to reduce the dimensionality of the space, to some selected number of dimensions. The authors assert that the dimensionality reduction step is crucial to the success of LSA, in that the process performs a powerful form of induction in order to capture generalisations about word usage. In addition, modelling performance is heavily dependent on the right dimensionality being chosen; this optimal space corresponding to "… the same dimensionality as the source that generates discourse, that is, the human speaker or writer's semantic space." (Landauer *et al*., 1998, p. 7).

Landauer and Dumais (1997) evaluated the psychological reality of their semantic distance measure (calculated as the cosine of the angle between SVD-reduced word vectors) by simulating the standardised synonym choice test taken by non-native English speakers who apply for admission to American universities. If LSA's representational space is effective in encoding the semantic similarity relationships between words, a test word should be 'closer' to the correct synonym than to any of the alternatives. LSA achieved 65% performance on the task, which was comparable to the typical foreign admission candidate. Landauer *et al.* (1998, p. 4) make the strong claim that "… LSA allows us to closely approximate human judgements of meaning similarity between words …" Although modelling of synonym test performance and other cognitive phenomena has been largely successful, Landauer *et al*. nevertheless have not yet demonstrated an explicit relationship between LSA's semantic distance measure and human judgements of semantic similarity.

In an ambitious research programme, Lund, Burgess and associates (*eg.* Burgess & Lund, *in press*; Lund & Burgess, 1996) have investigated the ability of semantic space models to account for a wide range of cognitive phenomena, such as lexical priming, deep dyslexia, syntactic parsing constraints, and decision making. Most of their work looks at the relationship between a geometric semantic distance measure (generally Euclidean distance) and behavioural data, though they also explore the role of the size and density of a word's 'semantic neighbourhood' in the representational space. Of particular interest are their simulations of semantic and associative priming effects (Lund, Burgess & Atchley 1995; Lund, Burgess & Audet, 1996), and the conclusions they draw from these simulations. We defer further discussion of this work until section 3.3.3.3 below.

Lund and Burgess (1996, Experiment 3) claim a linear relationship between their measure of semantic distance and (un/)primed lexical decision response times. They report correlation coefficients ranging from 0.25 to 0.35; the degree of association depended on the window size used and the precise form of the distance metric. However, their decision to use lexical decision latency, rather than the size of the priming effect, as the variable to correlate with semantic distance is questionable. There are certainly other factors which affect the lexical decision response time for a target word besides the semantic distance to its prime, which will obviously influence the correlation coefficient.

Other research exploring the psychological reality of co-occurrence statistics is reported by Levy and colleagues (Levy, Bullinaria & Patel, 1997; Patel, Bullinaria & Levy, 1998), who are primarily interested in the effect of varying model parameters on the quality of the resulting high-dimensional representations. In order to evaluate parameter settings, they use several sets of psychological data including published semantic category norms and word pairs judged to be near-synonyms. Results supported the 'semantic-ness' of their co-occurrence vectors. It is evident that their parameter exploration procedures also addressed the psychological validity of the various semantic distance measures examined, since modelling performance for the optimal parameter settings was always above the chance baseline. Interestingly, they have also managed to better the performance of Landauer and Dumais' LSA model on the same synonym test (achieving a score of 76%), using a window-based definition of co-occurrence to construct an (unreduced) 4000-dimensional semantic space from the BNC.

3.2 Validity investigation I: Semantic similarity judgements[1]

We have seen that people working with semantic space models have generally taken for granted the hypothesis that semantic distances are to some degree analogous to human intuitions of semantic similarity. Although the assumption of psychological reality is implicit in nearly all of the computational linguistics research using this class of model, it has not yet been directly investigated. The aim of the empirical work reported in this section is to address the *validity* issue: how *valid* is a semantic distance measure derived from the distribution of words in a large corpus? In order to establish the validity of any measuring instrument (such as semantic distance), its

[1] An earlier version of this section was reported in McDonald (1998a).

measurement data should be shown to co-vary with another, independent, source of data. Without the validation provided by a *criterion* measure grounded in psychology, semantic distance has no meaning outside the system in which it is measured. In order to assess the external validity of the corpus-derived measure, we employ semantic similarity judgements as the criterion measure. This validity investigation is presented in the larger context of an investigation into the representational basis of word-to-word similarity judgements.

The rest of this section is organised as follows. First, we briefly review the historical roots of the measurement of semantic similarity, and then introduce Miller and Charles' (1991) theoretical and empirical work addressing the basis of similarity judgements. Next, we define a semantic distance measure using the BNC-spoken subcorpus, and assess its validity using judgement data from Miller and Charles. Finally, we further evaluate the corpus-derived measure's predictive power on similarity ratings collected for two new sets of stimuli.

3.2.1 The measurement of meaning

The measurement of the semantic similarity between words has a long-standing place in experimental psychology. The work of Osgoode, Suci and Tannenbaum (1957) is an early example; here factor analysis and multidimensional scaling were applied to subjects' judgements of word meaning, measured on a variety of property scales. Since then, much work has been done in establishing the quantitative properties of the relatedness between words, and as a result the task of rating a pair of words for semantic similarity has achieved the status of an indisputable property of normal human ability. Semantic similarity is often treated as a random variable in experimental design, and is frequently taken into consideration when matching word stimuli.

The most common experimental methodology involves eliciting similarity judgements along an ordinal (n-point) scale. Ratings are averaged over subjects, yielding highly reliable measurements of semantic similarity. Similarity judgements are consistent over time; for example, the Pearson product-moment correlation coefficient between ratings of a set of 30 word pairs, made by two different groups of subjects 25 years apart (Rubenstein & Goodenough, 1965; Miller & Charles, 1991) is a remarkable 0.97 (p=0.01).

Goodman (1972) argues that similarity between entities cannot be established unless it is known in what *respects* the entities are to be judged. In the work of

Osgoode *et al*. (1957), the 'respects' are made salient to the subjects, in that the endpoints of the judgement scales are pre-determined (they are set to pairs of antonymous adjectives). It is clear that in a simple semantic similarity judgement task, subjects must determine their own 'respects' (or frame of reference) when making a comparison. Nevertheless, the robustness of the results seem to overwhelm any objections on these grounds. In an extensive review of the literature on similarity judgements, Medin, Goldstone and Gentner (1993) conclude that "… similarity is far from an empty concept with no explanatory power" (p. 275).

3.2.2 Contextual similarity

A foundational assumption of this thesis is that the meaning of a word, in some sense, is defined by its linguistic contexts of use. The central motivation for examining the relationship between semantic similarity and linguistic context stems from the view that one aspect of a word's cognitive representation is an amalgam of the contexts in which it occurs (Miller & Charles, 1991). In other words, a word's *contextual representation* is something distinct from other components of its mental representation (such as information contributed from phonological form and world knowledge) – it consists of knowledge of how that word is used.

Because it is possible to learn the meaning of a word from its linguistic surroundings only, the definition of contextual representation can be operationally restricted to exclude information from the extra-linguistic context. This restriction is in principle consistent with Cruse's (1986) observation that linguistic context often acts as a mediator between a word and its extra-linguistic context. Miller and Charles (1991, p. 8) express the relationship between contextual representation and word meaning in what they term the Strong Contextual Hypothesis:

> *Strong Contextual Hypothesis*: Two words are semantically similar to the extent that their contextual representations are similar.

Because of the observation that words from different languages (*eg.* <*department*, *Abteilung*>) or from different syntactic categories (*eg.* <*department*, *departmental*>) can be judged semantically similar, yet be found in completely different linguistic contexts, Miller and Charles (1991, p. 9) weaken their hypothesis:

> *Weak Contextual Hypothesis*: The similarity of the contextual representations of two words contributes to the semantic similarity of those two words.

This statement suggests that the similarity of the linguistic contexts in which two words occur should, to a certain degree, be informative about their semantic

similarity. Put another way, if two words can be substituted for one another in the same context without affecting plausibility, then these words are more often than not semantically similar. The Weak Contextual Hypothesis leads to a testable prediction: if the meaning of a word is closely tied to its contexts of use, then the similarity of meanings and the similarity of contexts should co-vary.

In order to test this prediction empirically, a means to estimate the similarity between two words' contexts of use – their *contextual similarity* – is required. Miller and Charles suggest two possible approaches. The first relies on *co-occurrence*: the set of words found in the immediate context of word w_1 and the set of words co-occurring with word w_2 are compiled (perhaps from a corpus), and a calculation based on the overlap in set membership can be construed as a measurement of the contextual similarity of w_1 and w_2. The second approach is based on the notion of *substitutability*: the degree that either of two words can plausibly appear in the context of the other reflects their contextual similarity.

Miller and Charles present an experimental substitution task called the 'method of sorting', which they use successfully to establish a measure of contextual similarity. Sets of sentences containing the *target* words (*eg.* <*monk*, *slave*>) were first extracted from the Brown corpus, and each target word was replaced by a dash '——'. Working with one pair of targets at a time, subjects decided for each sentence which target(s) could plausibly fit into the sentence context. Signal detection theory was used to compute the discriminability of contexts; contextual similarity (construed as the inverse of discriminability) was found to be linearly related to data collected from a semantic similarity judgement task. Their results thus confirm the Weak Contextual Hypothesis.

Rubenstein and Goodenough (1965) tried the co-occurrence-based approach: they calculated the contextual similarity between a pair of target words as a function of the number of words common to subject-generated contexts for each target word. Although word pairs with the largest amount of 'contextual overlap' also received the highest similarity ratings with this method, Miller and Charles argue that the substitutability approach is superior for estimating contextual similarity than an approach relying on co-occurrence, since Rubenstein and Goodenough's results indicated that co-occurrence information was not reliable for distinguishing the middle and lower ranges of the similarity scale. The method that Rubenstein and Goodenough used for calculating contextual similarity was quite primitive, however,

when compared to what can be achieved using the data-intensive techniques introduced in Chapter 2.

Experiment 1 was designed to test the validity of an objective distributional similarity measure, *Contextual Similarity*, derived from the co-occurrence information present in a large corpus of natural language. Semantic similarity ratings from Miller and Charles (1991) served as the criterion measure. We predicted that Contextual Similarity would correlate positively with rated semantic similarity, establishing the validity of the corpus-derived measure.

3.2.3 Experiment 1

3.2.3.1 Method

The lemmatised version of the 10M word BNC-spoken was used to construct a semantic space model. We created co-occurrence vectors for a subset[2] of the target words examined by Miller and Charles (1991) using a context window size of ±3 words, and each vector element encoded the value of the log-likelihood ratio statistic for the particular target-context word combination.

A statistically motivated procedure was sought for choosing the set of context words. The issue is one of reliability: since a target word is 'defined' by its co-occurrence with a set of context words in a certain corpus, the same set of context words should reliably represent the same target word even when the co-occurrence matrix is compiled from a different corpus. If co-occurrence vectors from two (or more) corpora can be shown to be similar, then we can be confident that the vectors are encoding a word's 'true' contextual behaviour. By the same logic, reliability of *context* words can be estimated, by comparing vectors of *target* words (corresponding to columns of the co-occurrence matrix). 446 contentive context words were selected using Kendall's coefficient of concordance *W* as a reliability statistic (for a detailed description of the procedure involved, see McDonald, 1997).

These parameter settings were employed in all of the semantic space modelling reported in this chapter and in Chapter 4 (*ie*. the semantic space was constant across simulations).

---

[2] Because of the unreliability of co-occurrence vectors created for very low frequency words (*cf*. section 2.4.3), we excluded pairs where one or both members had a BNC-spoken lexeme frequency of less than 25. This reduced the number of word pairs considered to 19, from the original set of 30 listed in Miller and Charles (1991, Table 1).

Table 3-1. Semantic and Contextual Similarity Measurements for 19 Target Word Pairs.

Target Word Pair	Mean Rating	Contextual Similarity
gem–jewel	3.84	0.278
boy–lad	3.76	0.746
coast–shore	3.70	0.194
midday–noon	3.42	0.384
furnace–stove	3.11	0.331
food–fruit	3.08	0.708
tool–implement	2.95	0.117
brother–monk	2.82	0.046
lad–brother	1.66	0.199
crane–implement	1.68	0.111
journey–car	1.16	0.104
cemetery–woodland	0.95	0.230
coast–hill	0.87	0.097
forest–graveyard	0.84	0.046
shore–woodland	0.63	0.129
monk–slave	0.55	0.046
coast–forest	0.42	0.099
glass–magician	0.11	0.045
noon–string	0.08	0.019

Note: mean semantic similarity ratings are from Miller and Charles (1991).

Next, Contextual Similarity was defined as the cosine of the angle between the vectors for each target word pair. Because the cosine measure is insensitive to vector length, it is useful for comparing words that differ in corpus frequency.

Finally, we created a co-occurrence measure similar to the one described by Rubenstein and Goodenough (1965) to use as a baseline measure of the similarity of contexts. It is necessary to establish that Contextual Similarity, which is derived from 'higher-order' co-occurrence information, is superior to a simple measure of 'contextual overlap', derived from 'local' co-occurrence statistics. The baseline measure was defined as the number of non-zero vector components shared by the members of a target word pair, divided by the smaller of the total non-zero components. To illustrate the calculation for the target pair <*food*, *fruit*>: 92 of the set of 446 context words appear within a ±3 word window of both *food* and *fruit* in the BNC-spoken; the vector for *food* has 269 non-zero elements and *fruit* has 108. For this pair, the contextual overlap is 92 / 108, or 0.852. If the semantic similarity between two words simply reflects the number of word types that co-occur with

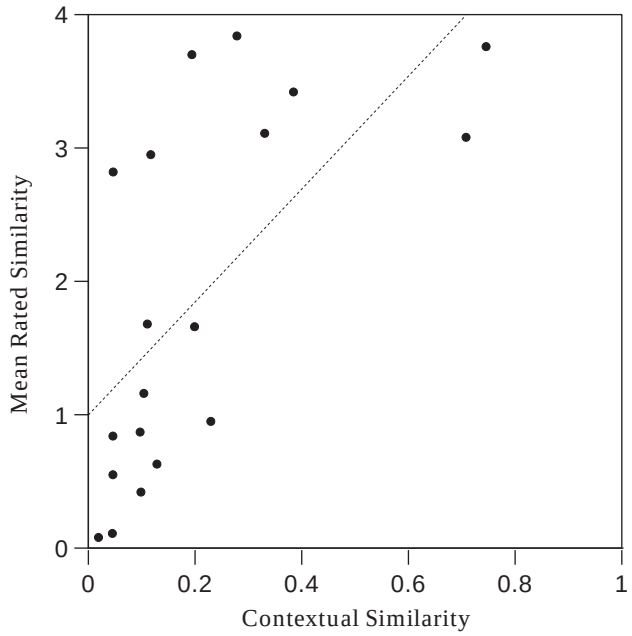

Figure 3-1. Semantic similarity plotted against corpus-derived Contextual Similarity (*n*=19).

both members of the pair, then the higher the rated similarity, the greater the overlap should be.

3.2.3.2 Results

Semantic similarity ratings from Miller and Charles (1991) and the corresponding Contextual Similarity values for 19 word pairs are given in Table 3-1. Figure 3-1 graphically displays the results of plotting Contextual Similarity against mean rated similarity. A linear relationship between the two measures was confirmed by a correlation analysis: r=0.65, 17 *df*, p<0.005, one-tailed.

There was no appreciable linear relationship between semantic similarity and the 'contextual overlap' baseline measure: r=-0.08.

3.2.3.3 Discussion

The moderate correlation obtained between the objective semantic distance measurements and the mean semantic similarity ratings establishes the validity of the corpus-derived Contextual Similarity measure: Contextual Similarity is significantly predictive of rated semantic similarity. The results of Experiment 1 also confirm Miller and Charles' Weak Contextual Hypothesis: to the extent that a co-occurrence vector is a useful model of a word's contextual representation, similarity

of the contextual representations for two words can be said to contribute to their semantic similarity. Because both the current results and those of Miller and Charles' sorting task support the Weak Contextual Hypothesis, even though obtained using completely different methodologies, an attractive underlying generalisation becomes apparent: the contextual representation of a word is formed from experience with that word in the linguistic environment.

Although wholly derived from co-occurrence counts, semantic space models can also be conceptualised as encoding substitutability: the more similar two words' co-occurrence vectors are, the more substitutable in context the two words should be. Note that in the model described here, it is possible for two words to have similar *semantic* contexts, even though they are members of different *syntactic* categories. Miller and Charles' (1991) reliance on a substitution task for estimating contextual similarity is deficient in this important respect: the words being compared are required to be of the same syntactic category. Measuring the contextual discriminability of morphologically related, but grammatically different word pairs such as <*department*, *departmental*> is not feasible using their sorting task; yet rating this pair of words for semantic relatedness is a task easily done by people. Moreover, measuring the Contextual Similarity between lexical representations in the current semantic space model is not subject to syntactic constraints,[3] since word vectors are simply compared as numerical entities. Therefore, a straightforward prediction is that the Contextual Similarity between words of different syntactic categories can be estimated in exactly the same fashion as in Experiment 1.

3.2.4 Experiment 2

A second experiment comparing human similarity judgements with Contextual Similarity was designed to investigate two issues. The purpose of Experiment 2A was to replicate Experiment 1, using newly-elicited similarity ratings for 30 pairs of *same*-syntactic category word pairs. It was also desirable that the stimuli be representative of several contentive syntactic categories, rather than be restricted to

[3] The inclusion of function words in the set of context words defining the dimensions of the semantic space (in conjunction with a narrow context window) would likely invalidate this statement. Using a similar approach to model construction, with one crucial exception being that function words were not excluded as component labels, Burgess and Lund (*in press*) find that their co-occurrence vectors also seem to encode the target word's syntactic category (see also Finch & Chater, 1992).

nouns (as in Miller & Charles, 1991), in order to assess the generality of the Contextual Similarity measure for other parts of speech.

Second, Experiment 2B used *different*-syntactic category target pairs as stimuli, in order to test the hypothesis that the model's vector representations primarily encode semantic, as opposed to grammatical category information. Words of dissimilar syntactic category, yet semantically related, should be more distributionally similar than semantically unrelated words.

3.2.4.1 Method

Subjects. Twenty-four questionnaires were distributed to members of the University of Edinburgh community who had volunteered to participate. Nineteen questionnaires were returned to the experimenter. All subjects were native English speakers.

Materials and Procedure. Semantic similarity judgements were collected using a ratings task in questionnaire format. A set of 60 pairs of target words was compiled, representing an intuitively broad similarity range. Four different randomisations of the materials were created, and half of the questionnaires presented the word pairs in reverse order, since Tversky (1977) has shown that similarity judgements can be asymmetric.[4] Because the Contextual Similarity measure is symmetric, this balancing was done in order to provide a corresponding symmetric measure of semantic similarity.

The 30 word pairs comprising Experiment 2A were of the *same* grammatical category; specifically 14 pairs of nouns, six of verbs, six of adjectives and four of adverbs. The 30 word pairs representing Experiment 2B were of *differing* syntactic category (*eg.* <*friend*, *social*>). These stimuli consisted of 14 noun-verb combinations, five each of noun-adjective and adjective-adverb pairs, and three each of noun-adverb and verb-adjective pairs. BNC-spoken lexeme frequency ranged from 40 occurrences per million to 1,189 / million.

Subjects were asked to rate "how similar in meaning" the words in each pair were, using a 9-point scale, where "a 9 represents a highly similar pair of words, and a 1 represents a pair of words that are completely unrelated in meaning." The instructions also encouraged participants to distinguish as many different degrees of

[4]Nine of the 19 returned questionnaires had the alternate presentation order.

Table 3-2. Materials for Experiment 2 and Mean Semantic Similarity Ratings.

Experiment 2A	Rating	Experiment 2B	Rating
divide-split	8.11	completely-total	7.42
awful-horrible	8.00	proposal-suggest	7.16
likely-probably	7.63	immediately-quick	6.58
beautiful-lovely	7.53	believe-opinion	6.37
various-different	6.89	financial-economy	5.58
discussion-conference	6.58	remind-memory	5.47
receive-accept	6.11	settlement-establish	5.21
food-bread	5.79	write-pen	5.00
action-performance	5.74	simple-clearly	5.00
normally-often	5.47	dinner-eat	5.00
consider-study	5.42	grow-life	4.68
officer-staff	5.00	friend-social	4.58
sea-river	4.63	interesting-attention	4.42
strong-heavy	4.21	information-tell	4.26
meat-body	4.21	basis-main	4.21
straight-easy	3.95	department-manage	3.05
respond-understand	3.95	possible-soon	3.00
story-reference	3.89	agreement-fairly	3.00
stupid-common	3.21	rich-enjoy	2.84
entirely-already	3.11	allow-health	2.32
door-hall	3.05	wear-warm	2.28
include-explain	2.95	special-definitely	2.26
metal-railway	2.89	recently-actual	2.26
provide-increase	2.84	effort-political	2.16
office-product	2.53	prepare-moment	2.00
almost-somewhere	2.53	early-create	1.74
thought-child	2.16	lose-truth	1.63
duty-method	2.11	income-involve	1.58
housing-music	1.26	slightly-husband	1.21
car-county	1.21	newspaper-continue	1.16

similarity as possible. Stimuli and their corresponding mean similarity ratings are given in Table 3-2.

Co-occurrence vectors for each word pair were extracted from the BNC-spoken, and Contextual Similarity and the baseline 'contextual overlap' measure were determined as in Experiment 1.

3.2.4.2 Results

Experiment 2A. A correlation analysis revealed a significant linear relationship between rated similarity and Contextual Similarity, for same-category stimuli: r=0.50, 28 *df*, p<0.005, one-tailed (see Figure 3-2). The correlation between mean semantic similarity and the baseline measure of contextual overlap was not significant: r=0.14.

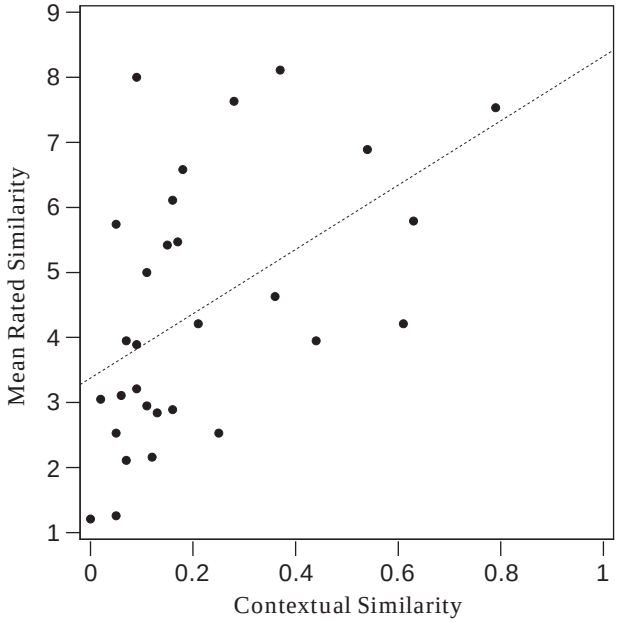

Figure 3-2. Semantic similarity ratings for same-category word pairs (Experiment 2A) plotted against their Contextual Similarity values (n=30).

Experiment 2B. The correlation between human similarity judgements and the model's measure of Contextual Similarity for cross-category word pairs was marginally significant: r=0.29, 28 *df*, p=0.061, one-tailed. There was no linear relationship between judged similarity and contextual overlap: r=-0.08.

3.2.4.3 Discussion

The results of Experiment 2A successfully replicated Experiment 1's findings, using a different (and larger) set of stimuli. A moderate correlation was obtained using materials chosen from four different grammatical categories. Although it did not reach statistical significance, the correlation between elicited similarity ratings and Contextual Similarity in Experiment 2B suggests that the semantic distance between the co-occurrence vectors for words belonging to different syntactic categories is also predictive of their rated semantic similarity. These results provide further support for the validity of the corpus-derived Contextual Similarity measure.[5]

[5] This is really the only sensible way to view the dependence between the two variables. The relationship between Contextual Similarity and semantic similarity is far from perfect; for instance, in Experiment 2A the Contextual Similarity measure accounted for approximately 25% of the rated semantic similarity variance. Obviously, the

The cross-category materials used in Experiment 2B may have given rise to a weaker correlation than the same-category stimuli because co-occurrence vectors contain a limited amount of syntactic information. Even though function words were excluded as context words during model construction, other grammatical categories can impose syntactic constraints. For example, the fact that adjectives tend to occur to a greater extent in the immediate context of nouns than that of verbs will help distinguish the vector representations of nouns from verbs.

3.2.5 General discussion

Experiments 1 and 2 explored the validity of an objective measure of semantic similarity derived from the distributional information present in a large corpus of spoken language. In Experiment 1, Contextual Similarity values were found to correlate significantly with published semantic similarity ratings. The correspondence between the data from the two measures confirmed the validity of the corpus-derived Contextual Similarity measure. The results of Experiment 2A offered further support for the validity of the measure, using a new set of materials. Experiment 2B showed that Contextual Similarity was (marginally significantly) predictive of elicited similarity ratings, even when constraints on the syntactic category of the target words were relaxed. The success of the co-occurrence-based procedure for predicting the semantic similarity of words differing in grammatical category overcomes one of Miller and Charles' (1991) motivations for replacing their Strong Contextual Hypothesis with the weaker version.

The present findings validate the assumptions of psychological reality made by computational linguistics-oriented semantic space research (*cf.* section 3.1.1): corpus-derived semantic distances do correspond (to a certain degree) to human intuitions of semantic similarity. The remarkable property of high-dimensional semantic space models is that they do not contain any *a priori* assumptions about psychological similarity or any encoded linguistic knowledge – they are constructed entirely from natural language output. The distributional characteristics of words appear to reflect, at least to a certain extent, their semantic properties.

There are several potential reasons why the correlations obtained between rated similarity and Contextual Similarity were non-optimal. First, there are variables affecting semantic similarity judgements which are simply not determinable from the

---

similarity of vector representations (or lack thereof) cannot be used as the basis for claims about the validity of human similarity judgements.

linguistic context: encyclopedic knowledge about the referents of the target words, for example (see also Tversky, 1977). Another influential factor is likely the nature of the vector representations themselves; a word vector 'smears' together the contexts for all appearances of the word in the corpus. Thus, *coast*, which is part-of-speech ambiguous between noun and verb, has only a single representation in the model. Distinct meanings corresponding to a single word form are similarly lumped together; for example *glass* can refer to either the drinking utensil or to the substance. The third potential reason for the non-optimal correlation is the *sparse data problem* that is pervasive in corpus-based modelling (*cf.* section 2.4.1). The corpus frequency of one member of a target word pair might differ substantially from the frequency of the other. This can result in a particular component of the vector for the lower frequency word encoding a value of zero simply because the corpus is too small – not because the context word never appears together with the target word in natural language. Even though the cosine measure compensates for frequency differences between members of a target word pair (since only vector direction is compared), a dimension value of zero is as informative as a dimension with a non-zero value.

In summary, the results of Experiments 1 and 2 provide support for Miller and Charles' claim that a word's contextual representation incorporates knowledge about its contexts of use. Given the assumption that a co-occurrence vector is a useful model of a word's contextual representation, the present results have confirmed the Weak Contextual Hypothesis, and additionally have demonstrated that the relative importance of contextual representations in judgements of semantic similarity can be quantified using simple distributional statistics collected from natural language output.

3.3 Validity investigation II: Semantic and associative priming[6]

In the second part of this chapter, we employ a different source of empirical data to evaluate the psychological plausibility of the information latent in co-occurrence statistics. Although Experiments 1 and 2 demonstrated a significant linear relationship between Contextual Similarity measurements and *off-line* similarity judgements, it is desirable to carry out further validation using *on-line* (timed) processing data. Since the central aim of this thesis is to show that the distributional characteristics of words are relevant to fundamental language processes, it is

[6] The research in this section was originally published as McDonald and Lowe (1998).

necessary to get a closer picture of the relationship between Contextual Similarity and on-line lexical processing behaviour.

The phenomenon of *semantic priming* is generally thought to reflect principles of lexical representation and organisation (Meyer & Schvaneveldt, 1971; for a review, see Neely, 1991). The widely-used semantic priming paradigm provides a minimal context – typically a single word – which allows close investigation of the factors believed to influence lexical processing. In general, the existence of a relationship in meaning between a *prime* word and a *target* word facilitates responses made to the target; this finding invites exploration of prime-target relatedness in terms of distance in semantic space. In this section we carry out a detailed reanalysis of an important recent semantic priming experiment (Moss, Ostrin, Tyler & Marslen-Wilson, 1995), in order to further establish the psychological validity of the Contextual Similarity measure, and thus provide further proof that co-occurrence statistics contain semantic information.

3.3.1 Lexical relations that support priming

Although more than 25 years of priming research has shown that the prior presentation of a related prime word tends to speed the recognition of a target word, the *type* of relation between the prime and target words necessary to produce the effect is still under dispute. The vast majority of semantic priming studies have concentrated on investigating words in a *taxonomic* relation. Materials typically consist of category coordinates, such as *<cat, dog>*; the conventional finding is that *dog* is responded to more rapidly and more accurately when preceded by *cat* than when preceded by an unrelated word such as *cap*.

Priming has been observed between words that are both semantically related and normatively associated[7] (*eg*. *<dish, plate>*; Moss *et al*., 1995), are semantically related only (*eg*. *<dance, skate>*; Fischler, 1977), or merely frequently co-occur in text (*eg*. *<hospital, baby>*; McKoon & Ratcliff, 1992). The presence of facilitation for semantically related words in the *absence* of an associative relation has been particularly controversial (McRae & Boisvert, 1996; Shelton & Martin, 1992). Moss *et al*. (1995) point out that *functionally* related words, where the referents are related in ways that can be described in non-taxonomic terms, such as the instrument pair

[7] The normative association strength between two words w_1 and w_2 is measured by the proportion of subjects who produce w_2 as the first word that comes to mind when presented with w_1 in a free association task.

<*hammer, nail*> or the script relation <*restaurant, wine*>, have often been assigned to the normatively associated but *non*-semantic condition in the experimental design, thus confounding semantic and associative relations.

Moss *et al*. address these issues in three priming experiments by orthogonally manipulating three factors: normative association (associated, nonassociated), type of semantic relationship (category coordinate, functional), and semantic relatedness (related, unrelated). Using auditory presentation and the lexical decision task, a priming effect was observed for both category coordinates and functionally-related items, leading Moss *et al*. to conclude that functional information *is* accessible during word recognition. Furthermore, they found reliable priming effects both with and without the presence of normative association. By showing that priming occurred for functionally related but nonassociated word pairs, Moss *et al*. uncovered a new source of information affecting the word recognition system. The second important finding from this set of experiments was the interaction obtained between the association and relatedness factors: the presence of normative association resulted in a significantly larger priming effect. This additive effect Moss *et al*. call the 'associative boost.'

3.3.2 Priming as distance in semantic space

As outlined in section 3.1.2 above, there has been considerable interest recently in the modelling of lexico-semantic phenomena using the distributional information contained in language corpora. The assumption that proximity in high-dimensional semantic space corresponds to semantic relatedness is becoming increasingly fruitful for psycholinguistic modelling (*eg*. Landauer & Dumais, 1997; Levy, Bullinaria & Patel, 1997; Lund, Burgess & Atchley, 1995). However, if corpus-based models are to provide an adequate explanation of lexical processing behaviour, they need to be able to replicate the full range of priming effects found with human subjects. Simulations should include the variety of lexical relations that have been shown to support priming, and also demonstrate the additive effect of normative association strength on response facilitation.

In order to address these concerns, we attempted to replicate each of the effects reported by Moss *et al*. (1995) using the corpus-derived Contextual Similarity measure introduced in section 3.2.3.1 above. In Experiment 3 (below) we present the semantic space model and examine the similarities and differences between human performance and the results using the Contextual Similarity measure. We then offer

an explanation for the additive effect of normative association, which is tested by corpus analysis in Experiment 4.

3.3.3 Experiment 3

If corpus-derived similarity measures are to account for semantic priming effects (as shown by Lund *et al*. [1995] for category coordinate stimuli), a crucial test of the approach will be to see how well they account for priming between functionally related items, as well as the 'associative boost' observed by Moss *et al*. These effects are the focus of Experiment 3, which aims to computationally replicate Moss *et al* 's Experiment 2 (a speeded auditory lexical decision task with single-word presentation of prime and target words).

3.3.3.1 Method

The design was identical to the original experiment, which varied three main factors: Association, Semantic Type and Relatedness. Subtype was nested under Semantic Type: half of the Category Coordinates were natural kinds (*eg*. <*dog, cat*>) and half were artifacts (*eg*. <*aeroplane, train*>). Correspondingly, the Functional semantic type was divided into words found in instrument relations (*eg*. <*knife, bread*>) and those in script relations (*eg*. <*circus, lion*>). Target words and their related primes were taken from Moss *et al*. (1995, Appendix 1).

Co-occurrence vectors for each of the stimuli were extracted from approximately ten million words of the written text portion of the BNC. Parameter settings were identical to those used in Experiments 1 and 2 above (±3 word window, the same 446 context words, vector components encoded the log-likelihood statistic).

Several of the stimuli turned out to have extremely low corpus frequencies (for instance, *mumps* and *cutlass* occurred three and four times, respectively, in the corpus) which meant that vectors created for these words would likely be unreliable (*cf*. section 2.4.3). Consequently, we excluded each prime-target pair that contained a word with a lexeme frequency of less than 25, and balanced the other conditions by removing their lowest frequency pairs, leaving 12 items in each cell.[8]

Next, we calculated the Contextual Similarity (realised as the cosine of the angle between vectors) between Related prime and target words. Contextual Similarity for

[8] This procedure meant that the target words in the Associated and Nonassociated conditions were no longer strictly matched for corpus frequency. However, the Contextual Similarity measure is relatively insensitive to frequency (but *cf*. section 2.3.6).

Table 3-3. Mean Contextual Similarity, Difference in Mean Contextual Similarity (Diff), and Amount of Priming in Milliseconds (Priming) for Prime-Target Pairs in each Condition.

	Associated				Nonassociated			
Condition	Rel	Unrel	Diff	Priming	Rel	Unrel	Diff	Priming
Cat Coord (all)	0.404	0.133	0.271	94	0.321	0.139	0.182	36
Natural	0.403	0.134	0.269	109	0.326	0.143	0.183	57
Artifact	0.405	0.131	0.274	78	0.316	0.136	0.180	16
Functional (all)	0.225	0.115	0.110	71	0.163	0.117	0.046	41
Script	0.230	0.118	0.112	80	0.164	0.090	0.074	31
Instrument	0.220	0.111	0.109	60	0.162	0.143	0.019	50

Note: Rel=related condition, Unrel=unrelated condition.

the Unrelated prime-target pairs was calculated as the mean Contextual Similarity of the target with each of the 11 other primes in the condition.

3.3.3.2 Results

We carried out a three-way analysis of variance (Association × Semantic Type × Relatedness) on the Contextual Similarity measurements, with the Natural and Artifact and the Instrument and Script subtypes collapsed into the Category Coordinate and Functional semantic types, respectively. Contextual Similarity values and the corresponding human response time data are summarised in Table 3-3.

The simulation results proved to be very similar to those found in the original experiment. There was a main effect of Relatedness, $F(1,92)=73.61$ $p<0.001$, indicating that collapsing over all conditions, semantically related prime-target pairs were more distributionally similar than unrelated prime-target combinations (mean Contextual Similarity values of 0.278 and 0.126, respectively).

We found a main effect of Semantic Type: Contextual Similarity was significantly higher for Category Coordinates than for items in the Functional condition, $F(1,92)=23.25$; $p<0.001$. There was also an interaction between Semantic Type and Relatedness, $F(1,92)=17.36$, $p<0.001$. From Table 3-3, it is clear that the Relatedness effect size is larger for Category Coordinates than for Functional items. These results differ from Moss *et al.* (1995), who found no reliable difference in the size of the priming effect between Category Coordinates and Functionally related pairs.

There was a significant interaction between Association and Relatedness, $F(1,92)=4.63$, $p<0.05$. The Relatedness effect was larger for Associated than for Nonassociated pairs; this interaction replicates the associative boost. ANOVAs on the separate Associated and Nonassociated conditions revealed significant

Relatedness effects, $F(1,46)=40.22$, $p<0.001$ and $F(1,46)=36.35$, $p<0.001$, respectively, which correspond to the human results.

Consistent with the original experiment, the simulation failed to show an interaction between Association and Semantic Type, $F(1,92)<1$, and there was no three-way interaction between Association, Semantic Type and Relatedness, $F(1,92)<1$.

Since the three-way ANOVA revealed a significant difference between the Category Coordinate and Functional semantic types, we ran ANOVAs on each condition separately, in order to examine the relationship between Contextual Similarity and the type of semantic relation more closely.

First, we carried out an ANOVA on the Functional materials. Contextual Similarity for Related word pairs was significantly larger than for Unrelated pairs, $F(1,44)=23.73$, $p<0.001$. We found no main effects of Subtype, $F(1,44)<1$, or Association, $F(1,44)=2.21$, $p=0.14$. There were also no reliable interactions between the combinations of factors: Association × Subtype, $F(1,44)<1$; Subtype × Relatedness, $F(1,44)<1$; Association × Relatedness × Subtype, $F(1,44)<1$. The interaction between Association and Relatedness was marginally significant, however, $F(1,44)=3.91$, $p=0.054$. A nearly identical pattern of results was obtained for the separated Category Coordinates.

3.3.3.3 Discussion

In summary, the pattern of results largely corresponded to those reported by Moss *et al.* The significant difference in Contextual Similarity between semantically related and unrelated prime-target pairs replicated the overall priming effect found with human subjects. The simulation also produced the associative boost – Contextual Similarity was higher for semantically related prime-target pairs that were also normatively associated than for nonassociated materials.

Functional relations *vs.* category coordinates

The main discrepancy between the simulation and original results was the significant interaction between Semantic Type and Relatedness obtained in the simulation. Although separate ANOVAs for each Semantic Type condition verified Relatedness effects for both Category Coordinates and Functional items, it is clear that Contextual Similarity between Category Coordinate targets and their related primes was higher than for Functional prime-target pairs.

It is worth considering why this interaction should occur in a model constructed from co-occurrence statistics. We suggest that this interaction is due to differences in the semantic roles typically filled by Category Coordinate and Functional items. For example, Contextual Similarity will be higher between *bread* and its Category Coordinates (such as *fruit* and *soup*) principally because Category Coordinates tend to occupy the same position in predicate-argument structure, *ie*. the patient role associated with verbs such as *serve* and *eat*. Category Coordinates are therefore highly substitutable in 'thematic' context. In contrast, items in Functional relationships tend to fill different semantic roles, such as instrument and patient, and will occupy different positions in predicate-argument structure. Moss *et al*.'s Instrument materials are a clear example of this; words were chosen to fit the template in (1):

(1) `you use a <prime> to do <target>.`

In fact, functionally related prime-target pairs almost never occur in this way in real text. However, the vector representation of each word is a superposition of many separate occurrences, so the vector reflects the template structure better than any individual context. Since the Contextual Similarity measure does not distinguish left from right context, functionally-related prime and target words will share less of each other's context, and therefore will be less substitutable in context than Category Coordinates. This would be sufficient to produce the Semantic Type × Relatedness interaction.

The associative boost

Although both Moss *et al*. (1995, Experiment 2) and the computational simulation demonstrated a Relatedness effect for both Associated and Nonassociated pairs, in a similar experiment, Shelton and Martin (1992) found no evidence of (human) facilitation for semantically related items that were *not* also normatively associated. Lund *et al.* (1995, Experiment 2) attempted to replicate this experiment using a similar corpus-based model. But their simulation results were also incompatible with those of Shelton and Martin, since they found a simulated priming effect for the Semantic-only materials.

Lund *et al*. suggested that Shelton and Martin's Semantic-only materials were in fact less semantically related than their Associated stimuli, pointing out that the semantic distances computed from their model for word pairs in the Associated

condition were smaller than for word pairs in the Semantic condition (although this difference only approached significance, p=0.061). As an explanation for Shelton and Martin's failure to find Semantic-only priming, this reasoning is problematic, since distances between word vectors are assumed to reflect semantic relatedness because, in part, of their accord with semantic priming data.

Lund *et al.* (1995, Experiment 3) investigated this discrepancy by using stimuli from Chiarello, Burgess, Richards and Pollock (1990), which were more carefully controlled for semantic similarity. In a simulation using these materials, they failed to find an interaction between Type of Relation (Associated, Semantic or Semantic+Associated) and Relatedness. A separate analysis of the Associated condition also failed to reveal a Relatedness effect. This result was in accord with data from human subjects (Lund *et al.*, 1995, Experiment 4), confirming their hypothesis that the existence of a semantic relationship between prime and target words was necessary to induce a priming effect. However, results from another (human) lexical decision experiment (Lund, Burgess & Audet, 1996, Experiment 1), using a new set of unrelated pairs, revealed a reliable effect in the Associated condition. Notably, there was no corresponding effect in their semantic space simulation using this new stimuli set.

The lack of an interaction between the Type of Relation and Relatedness factors in both of these priming simulations is inconsistent with the associative boost reported by Moss *et al.*, which was also found in our computational replication. These differences warrant further discussion.

Explaining the associative boost

Moss *et al.* suggest that the associative boost is due to priming at a different level of representation than the semantic level. They propose that associative priming is dependent on syntagmatic relationships between lexical forms, *ie.* associative relationships between words such *elbow* and *grease* develop because they frequently co-occur in language. This is consistent with neural network models of priming (*eg.* Moss, Hare, Day & Tyler, 1994; Plaut, 1995) that treat semantic and associative priming as due to fundamentally different types of information – associative priming effects are the result of contiguity between training items.

In contrast, we suggest that there is no need to treat semantic and associative relations differently, whether as separate levels of representation or distinct mechanisms. It is possible for the same corpus-based model to explain both

semantic priming and the additive effect of normative association. Specifically, the associative boost falls out naturally from the way that co-occurrence statistics were compiled in the present experiment.

In order to address the discrepancy between Lund *et al.*'s (1995, 1996) simulation results and ours regarding the presence of an additive effect of association, it is necessary to examine the models' parameter settings in more detail.

One important difference is the size of the context window within which co-occurrence frequencies are recorded. If Associated items frequently co-occur within the same window, their Contextual Similarity will be higher, on average, than matched Nonassociated pairs, simply because of their shared local context.[9] For example, if *cup* and *saucer* nearly always appear together within a small window in the corpus, their Contextual Similarity will be high compared with prime-target pairs which merely tend to occur in similar (but not overlapping) contexts. Furthermore, the larger the window size, the more overlap of the immediate context shared by prime and target, and hence the greater their Contextual Similarity. However, as the window size increases, the number of semantically irrelevant co-occurrences recorded for each word also grows, increasing the noisiness of the vectors, which also affects the Contextual Similarity measure, suggesting that there is an optimal window size for capturing this phenomenon. In order to verify this hypothesis, we constructed versions of the model where the window size varied between ±1 to ±5 words. The best performance with respect to the human data was achieved with a window of ±3 words.

Lund *et al.*'s (1995) claims about Shelton and Martin's materials can now be addressed. It may be true that several of the Associated pairs are more semantically related than items in the Semantic condition, though Shelton and Martin did attempt to control their materials for this variable through a relatedness judgement pre-test. But we suggest an alternative interpretation of Lund *et al.*'s observations: the Associated prime-target pairs are marginally closer in semantic space compared with the Semantic-only pairs because their co-occurrence vectors encode, in part, local co-occurrence behaviour as well as substitutability in context.

---

[9] Note that by 'shared context' we do not mean that the prime and target words are also members of the set of words labelling the dimensions of the space, but rather that context words will often be shared if the context windows around the prime and target overlap, thus contributing to vector similarity. For example, if *and* was a context word, then the co-occurrence frequencies *f*(*and*, *cup*) and *f*(*and*, *saucer*) would be incremented simultaneously with every occurrence of the sequence *cup and saucer* in the corpus.

Lund *et al.* (1995) may not have found a reliable interaction between the Association and Relatedness factors in their Experiment 2 for the same reason that no associative boost was evident in the results of two other simulations (Lund *et al.*, 1995, Experiment 3; Lund *et al.*, 1996, Experiment 1). The absence of this effect was likely due to the way the co-occurrence counts were collected. Although the window size was quite large (10 words), these simulations used the 200 most variant context words (*ie.* the columns of the co-occurrence matrix with the highest variance) as vector components. This set will mostly consist of function words, which we suggest are simply not sufficiently specific indicators of semantic context.[10]

Our prediction about the origin of the associative boost in our semantic space model can be easily tested: if the Contextual Similarity between two word vectors is affected by their local co-occurrence behaviour, we expect to find that the probability of local co-occurrence (in a three word window) to be greater for Associated Related pairs than for the Nonassociated Related pairs. This hypothesis is investigated in Experiment 4.

3.3.4 Experiment 4

Spence and Owens (1990) conducted a corpus-based investigation into the relationship between lexical co-occurrence frequency and normative association strength using the one million word Brown corpus. Their central finding was that associatively related word pairs tended to co-occur (within a window of 250 characters – approximately 50 words) significantly more often than pairs of words that were not normatively associated.

Experiment 4 tests a similar hypothesis, namely that the probability of local co-occurrence in a much smaller window is higher for Experiment 3's Associated Related materials than for the Nonassociated items. The null hypothesis is that the difference in co-occurrence probability between the Associated and Nonassociated Related word pairs is not distinguishable from chance.

3.3.4.1 Method

Although Moss *et al.*'s Associated and Nonassociated target words were originally matched for median frequency according to the Hofland and Johansson (1982)

[10] We tried a similar approach using the 200 most variant lexemes occurring within a ±3 word window as context words. In this case the Contextual Similarity differences between Associated and Nonassociated pairs were negligible.

counts, the same calculation using lexeme frequency in the 10M word corpus indicated that the median frequencies of the two groups were not equal (57 per million *vs.* 38 per million). This difference could bias a comparison of co-occurrence *frequencies* in favour of the Associated pairs; if the Associated targets occur more often in the corpus, they have more chance of co-occurring with their corresponding prime words. Therefore, rather than comparing raw co-occurrence counts, we estimated *conditional probabilities*, which allowed us to normalise for the frequency of the target word:

$$P(prime|target) = \frac{f(prime, target)}{f(target)}$$

3.3.4.2 Results and Discussion

A Mann-Whitney *U* test revealed a highly significant difference between the lexical co-occurrence probabilities for Associated and Nonassociated pairs (U=630, p<0.00001, one-tailed). The probability of an Associated Related prime co-occurring with its corresponding target was significantly higher than for the Nonassociated Related pairs. This suggests that differences in co-occurrence probabilities for word pairs in these two conditions may be responsible for the difference in Contextual Similarity measurements, because of the natural incorporation of local co-occurrence information into the vectors representations.

The results of Experiment 4 are consistent with Spence and Owens' (1990) finding that corpus co-occurrence frequency and normative association strength are correlated: high associative strength predicts frequent lexical co-occurrence. McKoon and Ratcliff (1992) have additionally provided evidence that word pairs with a high probability of local co-occurrence, but are not highly normatively associated (*eg.* <*hospital*, *baby*>) also give rise to priming. This result is also consistent with our account of priming, although it should be noted that McKoon and Ratcliff's materials were not controlled for semantic relatedness.

Since it appears that the associative boost in semantic priming can be attributed to the impact of local co-occurrence on vector representations, we hypothesise that the additive effect of normative association strength is better described as a variable subsumed by the more general phenomenon of local co-occurrence. Our model would therefore predict an additive effect of high co-occurrence probability on priming between words that are semantically related.

Lund *et al.* (1996) have argued that normative association strength and lexical co-occurrence frequency are only correlated for the cases where a semantic relation is also present. The results of Experiment 4 are consistent with this claim. Given that a semantic relationship holds between a prime-target pair, the prime word is significantly more likely to occur within a small window of the target word if the prime and target are also normatively associated.

3.3.5 General discussion

Experiments 3 and 4 have demonstrated that a semantic space model is capable of explaining the priming effects obtained with materials representing a broad range of lexical relations. The corpus-derived measure of two words' distributional similarity, Contextual Similarity, corresponded well to the pattern of lexical decision facilitation observed in human subjects.

The computational replication of Moss *et al.* (1995) made four important contributions. First, Experiment 3 demonstrated that functional information is accessible directly from the linguistic environment. Functional relations are often considered to be represented as extra-linguistic, schema-based or episodic knowledge – this information was revealed using the same representational model that positions category coordinates close together. Although functional relations are implicit in simple distributional statistics, it appears that the co-occurrence patterns of category coordinates and functionally related words are not equally informative. This was indicated by the interaction between Semantic Type and Relatedness in the simulation, which was not observed in the original experiment. We presented an explanation for this effect based on the idea that co-occurrence vectors encode thematic constraints as well as semantic regularities. Specifically, we suggested that Category Coordinates typically fill the same semantic roles at the level of predicate-argument structure, whereas Functional items tend to occupy different roles.

The second contribution concerns the additive effect of normative association strength on the basic semantic priming effect observed by Moss *et al.* The interaction between Relatedness and Association (or 'associative boost') was modelled using the corpus-derived measure. Both lexical decision facilitation and Contextual Similarity between prime and target were greater for semantically related prime-target pairs that were also normatively associated, compared with pairs that were semantically related only. We offered an explanation for why the associative boost occurs in the corpus-based model, and why previous research has failed to find this

effect. We argued that the methodology used to collect co-occurrence statistics has a substantial impact on the 'associative' properties of the resulting vector representations; specifically, associated word pairs are those that have a high probability of local co-occurrence. This hypothesis was confirmed in Experiment 4.

Third, the simulation has addressed a controversial point in the priming literature: the independent effects of association strength and semantic relatedness have been offered as evidence for distinct, qualitatively different priming mechanisms or levels of representation. The results of Experiments 3 and 4 question the need for this distinction. A single level of representation can capture a wide range of lexical relations that support priming.

Finally, and most importantly, the psychological validity of a representational model built from co-occurrence statistics has received further support. The on-line processing differences revealed by the semantic priming paradigm were largely captured by the information contained in the corpus-based semantic space model. The measure of distributional similarity we used to compare vector representations was sensitive to two types of semantic relations – category coordinates and functional relations – as well as the additive effect of association strength. To the extent that the phenomenon of semantic priming draws upon representations of word meaning, we can conclude that co-occurrence vectors provide a valid and objective medium for representing important aspects of word meaning.

A caveat is necessary for any claim made about the role of semantic space models in explanations of semantic priming: it is clear that a high degree of vector similarity is not a *sufficient* predictor of (human) facilitation; if so we would expect to invariably observe priming effects between a target word and its closest 'neighbours' in semantic space. This hypothesis is intuitively false, as examination of typical nearest neighbour lists indicates. For example, Lund and Burgess (1996) list <*lace*, *pink*, *cream*, *purple*, *soft*> as the five nearest neighbours of *lipstick* and <*beauty*, *prime*, *grand*, *former*, *rolling*> for *triumph*. These are typical examples of nearest neighbours obtained using this type of model, yet it is extremely doubtful that human processing facilitation would be found between the target and even its closest neighbour in these examples.

3.4 Summary

The main contribution of this chapter was to present important new support for the psychological reality of the distributional information inherent in the linguistic environment. After a brief review of the small body of previous work in the area, the semantic nature of co-occurrence vector representations was established through two case studies. We began by summarising research in both language engineering and cognitive modelling which was founded on the underlying assumption that distributional similarity reflects semantic relatedness. Experiments 1 and 2 demonstrated positive correlations between the corpus-derived measure, Contextual Similarity, and two sets of elicited similarity ratings. In Experiments 3 and 4, the same measure was used in a computational reanalysis of an influential lexical priming experiment. Besides showing that a number of the lexical relations that support priming are implicit in simple distributional statistics, the results of this simulation offered a compelling solution to a controversial issue in the priming literature. There appears to be no reason to postulate more than a single level of representation in order to explain the additive effect of association strength on semantic priming.

4.
Representing Context-dependent Meaning

Chapter 3 assessed the validity of a representational model derived from distributional statistics in semantic processing situations where a target word is encountered in a minimal, single-word context. However, this is an unnatural situation in real language use. A word is most often found in a connected linguistic context – its meaning being to a large extent determined by that context. As well as influencing semantic interpretation, the particular context in which a word occurs affects measures of processing difficulty, such as response times. In this chapter, we explore the context-dependent nature of lexical processing and interpretation, and critically assess how a corpus-based representational model can account for the behavioural evidence. Although largely successful, these investigations nevertheless illuminate several potential deficiencies of the representational model. Before examining the psychological evidence for semantic context effects, we begin by looking at their description from the linguistic point of view.

4.1 Semantic context and interpretation

4.1.1 The linguistic view of meaning variation

A linguistic theory of lexical semantics must address the problem of how word meaning interacts with (or alternatively, is dependent on) the local linguistic context. The semantic contribution of a particular word to sentence and discourse meaning can vary widely, depending on the precise context in which it appears:

> There seems little doubt that such variation is the rule rather than the exception: the meaning of any word form is in some sense different in every distinct context in which it occurs (Cruse, 1986, p. 51).

To illustrate the notion of the context-dependent variation of word meaning, below are three instances of *firm* retrieved from the BNC:

(1) a. and you have to be *firm* and not let anyone in.
b. see the availability of rooms and then we'll make a *firm* decision.
c. sales were up by nineteen percent and there's also *firm* evidence in these results of margin improvements …

One can identify distinct, yet related meanings for each token of *firm* in (1a-c), most easily by labelling each instance with a synonymous expression. We might come up with STRICT for (1a), FINAL for (1b), and SOLID for (1c). It is clear that interpretation of *firm* in each case depends on the context.

Cruse (1986) identifies two principal ways that context interacts with word meaning. The first, sense *selection*, is illustrated by (1a-c). Sense selection, where the context selects one sense from a set of discrete meanings associated with a word, is in turn distinguished from sense *modulation*, where the context modulates the meaning of a single sense. For example, in (2a) the weight of the car is highlighted by the context, and in (2b) its price:

(2) a. The car crushed Arthur's foot. (Cruse, 1986, p. 53)
b. We can't afford that car.

Cruse also distinguishes these two types of context-dependent meaning variation in terms of discreteness: sense selection is characterised by discrete steps along some meaning scale, whereas the nature of sense modulation is continuous. From the psychological point of view, meaning variation can be described as either discrete or continuous. The task used to investigate context-dependent differences in meaning of the same word form also determines the type of scale. A task such as categorising corpus citations according to shared meaning (Jorgensen, 1990) is naturally compatible with the discrete view of meaning variation; in contrast, having human judges rate the inter-relatedness of the different 'meanings' of an ambiguous word (Durkin & Manning, 1989) yields a continuous measure of meaning variation. Thus, linguistically-motivated distinctions between contextual selection and modulation are not necessarily relevant for psychological measurement. We propose that the context-dependent variation in meaning between two tokens of a particular word

should be measurable just as semantic similarity is measurable between two different words. It is clear that any attempt to represent lexical semantic information (from either the linguistic or psychological viewpoint) needs to adequately deal with contextual variation.

4.1.2 Meaning variation and vector representations

The relevance of the above synopsis of the context-dependence of word meaning to the current thesis should now be apparent: the simple definition of a word as a unique orthographic type is problematic for the semantic space approach to representing meaning. A word's vector representation is a conflated collection of co-occurrence data from every use of the word in the corpus; it is ignorant of the variations in meaning attributable to the individual contexts in which it occurs. A word vector is a superimposition of its individual occurrences, 'smearing' together all degrees of contextual variation (conventionally described as lexical ambiguity, polysemy or homonymy).

Returning to the corpus examples of *firm* in (1a-c) above, it is clear that the meaning of *firm* is different in each of the three contexts; consequently judgements of the semantic similarity between *firm* and *eg*. **strict**[1] should also depend on the context. Intuitively, the rated similarity for this pair of words would be higher when *firm* is presented in context (1a) than if presented in (1b) or (1c). Obviously, a standard vector similarity or distance measure applied to the co-occurrence representations for *firm* and **strict** will not capture the influence of context. So what sort of information is needed to model context-dependent semantic similarity? This information is necessarily some combination of the meaning conveyed by the linguistic context in which the word occurs, in conjunction with extralinguistic sources (*eg*. encyclopaedic or world knowledge). However, the only other information available in a semantic space model is that implicit in the co-occurrence representations of the other words in the context. If the aspects of meaning that are *relevant* to the context-specific meaning of *firm* could be determined, then a corpus-derived similarity measure that weights these relevant semantic properties higher than irrelevant properties would compute different values for the *firm*-**strict** relationship in each context, yielding estimates of context-dependent semantic similarity.

[1] Here, we use **boldface** to designate a word standing for a particular 'meaning' of an ambiguous *word*.

The distributional characteristics of the words in the immediate context of the ambiguous item are a potential source of information about relevance. It seems likely that the words in a semantically coherent context would occupy a narrow region (or form a *cluster*) in the high-dimensional representational space; the closer in meaning these context words are to each other, the tighter the cluster. The aspects of meaning common to the words in the context could thus be identified with the subspace where the words are more tightly clustered – and 'tightness' could be determined using standard statistical measures of variability.

In Experiment 5, we develop a context-dependent similarity measure according to this distributional view of relevance. The measure is based on the 'adaptive scaling' algorithm developed by Kozima and Ito (1995). Their method involves first constructing a high-dimensional semantic space from a machine-readable dictionary. Next, they apply a vector distance measure in order to rank words in the lexicon according to their semantic distance from a particular set of context words. For example, the closest words to the context set {BUS, CAR, RAILWAY} were words having to do with transport, such as *motor* and *road*. In contrast, the closest words to the context word set {BUS, SCENERY, TOUR} were concerned with tourism, such as *abroad* and *tourist*. Kozima and Ito's distance measure was affected by the distributional properties of the words in the context set; dimensions with high variability were weighted less because these dimensions were assumed to be less relevant to the aspects of meaning common to the words in the context set. Although Kozima and Ito used their adaptive scaling method to rank the words in the lexicon by their weighted semantic distance to a *set* of words, it is apparent that their procedure could be easily modified to estimate the context-dependent semantic similarity between *single* words, in particular between an ambiguous word and a word standing for one of its meanings (*eg*. *firm*-**strict**). The variability of the positions of the context words along each spatial dimension could be taken into account by a weighted semantic distance measure.

To illustrate the anticipated effect of the weighting procedure, consider the hypothetical context word distributions in Figure 4-1 for the two instances of *firm* in (1a) and (1c). Notice that the dispersion of the words in the STRICT context (1a) along Dimension 1 is more pronounced than their dispersion along Dimension 2. Under the above definition of relevance as the inverse of variability, Dimension 1 would be considered less relevant than Dimension 2, and consequently any difference between the positions of *firm* and **strict** along this dimension should be

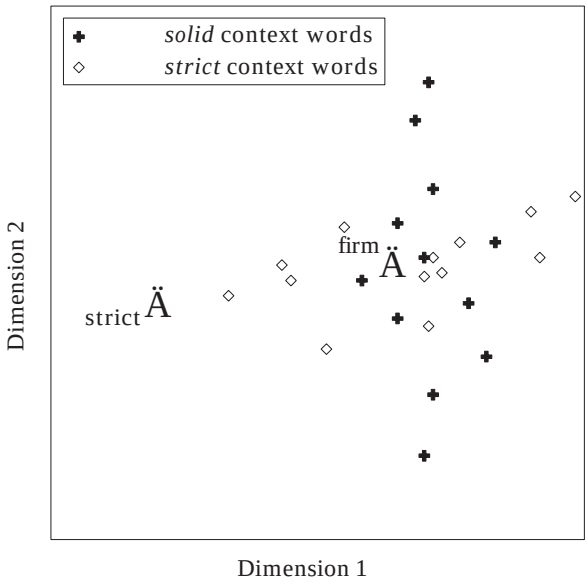

Figure 4-1. Hypothetical context word distributions in a 2-dimensional subspace for two different corpus occurrences of *firm*.

de-emphasised (weighted less by the semantic distance measure). In contrast, Dimension 1 is less variable – and therefore more relevant – for the SOLID context; the difference between the positions of *firm* and **strict** along Dimension 1 needs to be preserved. So, as a result of the weighting procedure, the semantic distance between *firm* and **strict** would be smaller when *firm* occurs in the STRICT context (1a) than when *firm* occurs in the SOLID context (1c), which corresponds to our intuitions. By implementing this procedure, a corpus-derived semantic distance measure can now be made dependent on the local linguistic context.

Cruse's (1986) work from the lexical semantic perspective on how the meaning of a word varies depending on its context receives some support from the psychological literature. For instance, Barsalou (1982) showed that concept similarity judgements vary according to the particular context. For example, in a PETS context, *snake* and *raccoon* were judged to be more similar than if no context was provided. Does Barsalou's result also hold with words, when represented in terms of their contexts of use? A testable hypothesis is whether the relevance-weighting procedure can realistically influence the semantic distance between an 'ambiguous' word and a word standing for its contextually-appropriate meaning (the *proxy*). To illustrate, the proxy word **price** intuitively should prove to be more similar to *car* in context (2b) than in (2a), because in (2b) the context highlights this

aspect of the meaning of *car*. The following computational pilot experiment tested this intuition using the same representational model presented in Chapter 3, in conjunction with a relevance-weighted version of the Contextual Similarity measure.

4.1.3 Experiment 5

4.1.3.1 Method

In order to test the hypothesis that context-dependent similarity can be estimated using the proposed relevance-weighting procedure, a suitable set of materials would ideally consist of ambiguous words each presented in more than one linguistic context – each context appropriate for a different 'meaning'. As a proof of concept, we decided to examine ambiguous items in short two-word contexts consisting of words normatively associated with one of the word's 'meanings'. If no contextual appropriateness effect is apparent using these contrived contexts, then it is unlikely to be found using more naturalistic materials.

Materials and Design

The first step was to compile a list of ambiguous stimuli, by exploiting Durkin and Manning's (1989) norms for the production frequencies of the various meanings of polysemous and homonymous words. In their norming study, participants were directed to produce the first meaning that came to mind when presented with the ambiguous word in isolation. The range of definitions (single words or phrases) produced were assumed to represent the range of senses or meanings associated with that word.

We took 20 ambiguous words from these norms, subject to the constraint that each word had exactly two meanings with production frequencies of 10% or higher. For example, for the item *orange*, subjects produced the definition FRUIT 74% of the time and COLOUR 26% of the time. In order to choose a word to stand for the dominant meaning of the ambiguous item (the Related proxy), we retrieved its highest-strength normed associate from the on-line version of the Edinburgh Associative Thesaurus[2] (EAT; Kiss, Armstrong, Milroy & Piper, 1973). The Appropriate context was created by embedding the ambiguous word in a two-word context, where the context words were biased indicators of its dominant meaning (strong associates of the ambiguous item from the EAT). Inappropriate contexts

[2] http://www.cis.rl.ac.uk/proj/psych/eat.html

Table 4-1. Design of Experiment 5 with Example Materials.

Context		Ambiguous Item	Type of Relation Related	Control
Appropriate	apple, fruit	orange	**lemon**	aside
Inappropriate	colour, red	orange	**lemon**	aside

were created using associates from the EAT that were deemed relevant to the non-dominant meaning.[3] Finally, control words matching the related proxy words in lexeme frequency and length were randomly selected from the BNC-spoken. Example stimuli are listed in Table 4-1, and the full set of materials can be found in Appendix C.

The experimental design was therefore 2 × 2: Context (Appropriate, Inappropriate) and Type of Relation (Related, Control). The Control condition was included in order to ensure that any observed Context effect could not be attributed to an unanticipated confounding variable, such as frequency differences between the Appropriate and the Inappropriate context words. An interaction between the two factors was predicted: there should be a Context effect for the Related proxies (*eg*. **lemon** should be more similar to *orange* in the FRUIT context of *apple* and *fruit* than in the COLOUR context of *colour* and *red*) but no effect for the Control words (*eg*. **aside** should be approximately equally similar to *orange* in both contexts).

Procedure

Vectors for all words were extracted from the BNC-spoken using the same procedure and parameter settings employed in Chapter 3 (±3 word window, same 446 context words, vector components encoded the log-likelihood statistic). Relevance-weighted Contextual Similarity between each ambiguous item (in either the Appropriate or Inappropriate context) and its corresponding Related and Control words was calculated as follows. First, the relevance r_i for each dimension i of the representational space was defined as the ratio of the standard deviation s_i of the positions of all words in the context C (which includes the ambiguous item) along dimension i, over the maximum standard deviation s_{max} for C:

$$r_i = \frac{s_i}{s_{\max}}$$

[3] For ten items, the Appropriate and Inappropriate contexts also selected different parts of speech; *eg*. *sink* as noun and *sink* as verb.

For each target *t* (either the Related proxy or the unrelated Control) the vector representing the ambiguous word *a* is moved to a new position in the space *a'* according to a function of *r* and its current distance from *t*:

$$a_i' = a_i + r_i(t_i - a_i)$$

If *r* is large, then any difference in the value of component *i* between *t* and *a* is made less prominent than if *r* is small. Finally, weighted Contextual Similarity is calculated as the cosine of the angle between *a'* and target *t*.

Note that because the algorithm moves the vector for the ambiguous word (rather than incorporating the relevance weights directly into the cosine expression), it would be straightforward to substitute another semantic distance measure.

4.1.3.2 Results and Discussion

An ANOVA revealed a main effect of Type of Relation, $F(1,19)=11.29$, $p<0.01$, as well as a main effect of Context: $F(1,19)=6.60$, $p<0.05$. As anticipated, the ambiguous items were significantly more similar to their meaning proxies than to frequency-matched control words. This result was qualified by the interaction

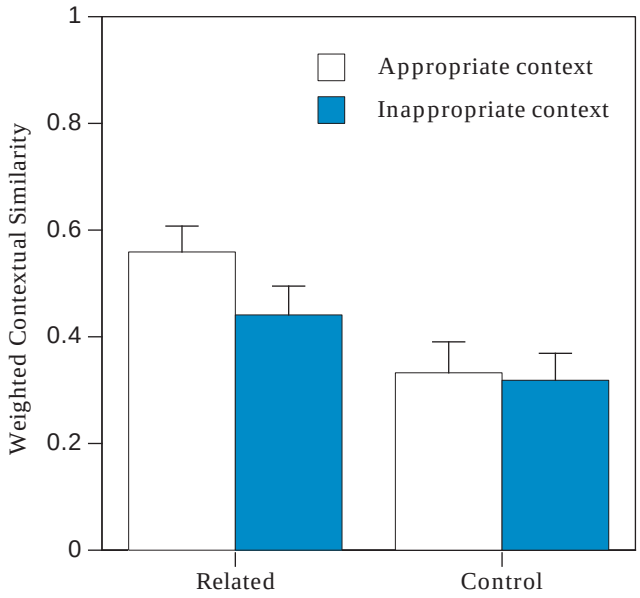

Figure 4-2. Weighted Contextual Similarity as a function of type of relation and contextual appropriateness (bars indicate standard error).

obtained between Context and Type of Relation: $F(1,19)=6.17$, $p<0.05$. Weighted Contextual Similarity between ambiguous items and their Related targets was significantly higher in the Appropriate context than in the Inappropriate context, whereas there was no such Context effect for the Control targets (see Figure 4-2). To illustrate this result using the above example: when the similarity measure was weighted according to dimension relevance, **lemon** was more similar to *orange* in the FRUIT context than in the COLOUR context, and there was only a negligible difference in relevance-weighted Contextual Similarity between the control word **aside** and *orange* between the two contexts, which confirmed experimenter intuitions. Without the weighting procedure, the Contextual Similarity between the vectors for **lemon** and *orange* would be identical in both the Appropriate and Inappropriate contexts.

Landauer and Dumais (1997) investigated a related semantic space approach in order to model human homograph disambiguation. Using a published set of sentence priming materials, they were able to simulate the context-dependent pattern of lexical decision facilitation observed with human participants. The final word of each sentence prime was a homograph, and responses were made to target words considered to be related to the distinct meanings of the homograph. Landauer and Dumais' approach to modelling this task was to average together the vector representations for all content words in the context before the homograph, and the resulting *centroid* vector was indirectly evaluated for its ability to capture the meaning of the context by measuring the similarity between the centroid vector and the vectors for words standing for the different meanings of the homograph. An example item was "Thinking of the amount of garlic in his dinner, the guest asked for a *mint*", with **candy** as the target word appropriate to the meaning of *mint* in this context, and **money** as the inappropriate target. Contextually appropriate targets were facilitated according to both human response time data and Landauer and Dumais' vector similarity measure, indicating that the averaged-vector representation of the sentence context contained sufficient information to simulate the priming effect.

Although Landauer and Dumais' approach was successful in capturing the pattern of human behaviour, it differs from our approach in terms of the assumptions made about the way sentence priming operates. The method of estimating context-dependent similarity described here implements the view that priming is essentially a word-to-word phenomenon, which is best simulated using a measure of the representational similarity between two word vectors. In contrast,

Landauer and Dumais model priming as the similarity of a word and an averaged 'context' vector. It remains to be seen if there is any empirical motivation for preferring one method over the other (*cf.* section 4.3.3).

We believe that our procedure of weighting the dimensions of the representational space according to relevance is consistent with Barsalou's (1982) demonstration of context-dependent similarity effects, while circumventing the 'problem' of representing a range of meaning variation by a single co-occurrence vector. However, the current approach to operationalising relevance is limited in one important respect – *only* linguistic context is used, and then only numerical representations of the words in the context. As Cruse (1986) notes: "It must not be forgotten, of course, that contextual relevance goes beyond the purely linguistic context and embraces the whole context of situation" (p. 101). And McNamara and Miller (1989) point out that determining contextual relevance is subject to individual differences in comprehension: "Undoubtedly, the salience of many features depends greatly on context and on the personal history of the perceiver" (p. 366).

4.2 Contextual constraint

We now turn to a related area of research on the effects of semantic context, this time concentrating on the psychological perspective. *Contextual constraint* refers to the influence of the local linguistic context on the predictability of a target word presented in that context. In other words, the more strongly a target word is constrained by the context, the greater the likelihood that the target will be observed. The amount of constraint in the relationship between a target word and a context can be estimated using subjective measures such as production (or *cloze*) probability in the sentence completion task (*eg.* Schwanenflugel, 1986).

The basic behavioural finding from research on contextual constraint is known as the *semantic congruity effect*. The typical experimental paradigm involves visual presentation of an incomplete sentence context, followed by a lexical decision made to a subsequent completion word (*eg.* Schwanenflugel & Shoben, 1985), but the congruity effect has also been observed in measurements of event-related brain potentials (ERPs) during auditory presentation (Van Petten, Coulson, Rubin, Plante & Parks, 1999). Word recognition is facilitated when the completion word is semantically congruous with the sentence context (3a), compared with a completion word that is in either a neutral (3b) or incongruous relation (3c) to the context.

Manipulating either the target word or the context can result give rise to the congruity effect.

(3) a. The woman took the warm cake out of the *oven.*
b. The woman took the warm cake out of the *car.*
c. The woman took the warm cake out of the *fun.*

The *amount* of constraint (or constraint *strength*) exerted by the linguistic context is another variable of interest. This quantity is standardly equated with the predictability of the most expected target word in the context, which is measured by its cloze probability – the proportion of subjects who produce this word when given a sentence completion questionnaire. Constraint strength has also been described as the degree to which the context constrains expectations about the identity of the missing word. For example, the context in (3a) is highly constraining, since *oven* was produced as a completion by 98% of the subjects in Schwanenflugel's (1986) norming study. In contrast, the incomplete sentence context (4) is low in constraint strength, since the completion with the highest production probability (*accident*) was offered by only 25% of participants.

(4) Everyone gathered around to look at the ____.

The processing behaviour of sentence completions *other* than the most expected is also potentially interesting. For instance, a constraining context might give rise to different effects for two different types of target words; compare (5a) where *beach* is both expected and semantically congruous (with a cloze probability of 0.78 according to Schwanenflugel's norms), with (5b), where *lake* is unexpected though still semantically congruous (with cloze probability of 0.13).

(5) a. On a hot summer's day many people go to the *beach.*
b. On a hot summer's day many people go to the *lake.*

Although the finding of lexical processing facilitation for target words of type (5a) is relatively uncontroversial, mixed results have been obtained for targets of type (5b). In their ERP study, Van Petten *et al.* (1999) found a small, but reliable effect of cloze probability. The N400 effect (an indicator of semantic incompatibility) was larger for low-cloze completions. Schwanenflugel and Shoben (1985) propose that cloze

probability and constraint strength interact, with unexpected yet congruous targets being facilitated in low-constraint strength contexts only.

4.2.1 The Feature Restriction Model

Contextual constraint is generally assumed to apply at the semantic (or conceptual) level of processing, and its effect is explained by appealing to semantic feature theories of meaning representation.

The Feature Restriction Model (Schwanenflugel & LaCount, 1988) is an attempt to explain the effects of manipulating contextual constraint (either constraint strength or the semantic congruity of the target word with preceding context) on lexical processing effort. This model assumes that context imposes semantic feature restrictions on upcoming words; depending on the strength of the context, these restrictions can range from narrow (or detailed) to broad (or general). If the upcoming word satisfies these feature restrictions (*ie*. it is semantically congruous with the context), then processing is facilitated, resulting in the basic congruity effect.

Schwanenflugel and LaCount (1988) argue that readers generate more detailed feature restrictions for upcoming words in highly constraining contexts than in less constraining contexts; consequently fewer words will meet these restrictions in the former case. They found that lexical decisions made to target words semantically related to the most expected completion were facilitated only when preceded by a low-constraint sentence context (compared with a neutral context condition). Their finding that words semantically related to the expected completion were not facilitated when the context was highly constraining indicates that a large amount of feature overlap (as semantic relatedness is typically construed in this model) was not sufficient; the target word was required to meet *all* the feature restrictions imposed by a highly constraining context in order for processing to be speeded.

4.2.2 The Contextual Relevance Model

Certain predictions of the Feature Restriction Model regarding contextual constraint effects could conceivably also be captured by the model of lexico-semantic representation presented in Chapters 2 and 3, in conjunction with the dimension-relevance weighting mechanism proposed in section 4.1.3.1 above. Semantic feature restrictions could alternatively be described as the relevant (or salient) aspects of

meaning evoked by the linguistic context. It is clear that expressing these aspects of meaning as feature restrictions often seems redundant with the lexical items in the context: for example, the restrictions imposed by the sentence context (5) (repeated below) would be captured by something like [PLACE VISITED BY PEOPLE], [PLACE VISITED DURING HOT WEATHER], [PLACE VISITED IN SUMMER]. It would be desirable to express such feature restrictions *objectively*; this is provided by a method related to the technique for estimating context-dependent semantic similarity developed in section 4.1.3.1.

(5) On a hot summer's day many people go to the ____.

We propose that the semantic congruity of a sentence completion with its preceding context can be estimated as a function of the semantic distance between the completion word and each content word in the sentential context. Furthermore, we believe that the semantic distance measure should take into account the variability along each dimension, and should weight the dimensions accordingly. For instance, in (5) the dimensions of the representational space that vary the least with respect to the vectors for the content words *hot*, *summer*, *day*, *people* and *go* should contribute greater weight to the vector similarity measure than dimensions whose variance is larger. Even if sentence context (5) was equally similar to the vectors for *beach* and an incongruous completion *asset* (according to an unweighted proximity measure), the relevance-weighted measure should be able to capture this difference in semantic congruity between the context with *beach* and the context with *asset*. We shall refer to this alternative approach to characterising contextual constraint as the Contextual Relevance Model.

Put another way, the Contextual Relevance Model describes the relationship between a constraining sentence context and a target word as the relevance-weighted function of the proximity of the target word vector to the *region* of semantic space occupied by the words in the sentence context. Thus, the Feature Restriction Model can be replaced with a model where contextual constraint effects on lexical processing are simulated using only the distributional information contained in a large corpus of natural language output. In the Contextual Relevance Model, semantic feature restrictions are operationally equated with the representational subspace that is relevant to the current linguistic context.

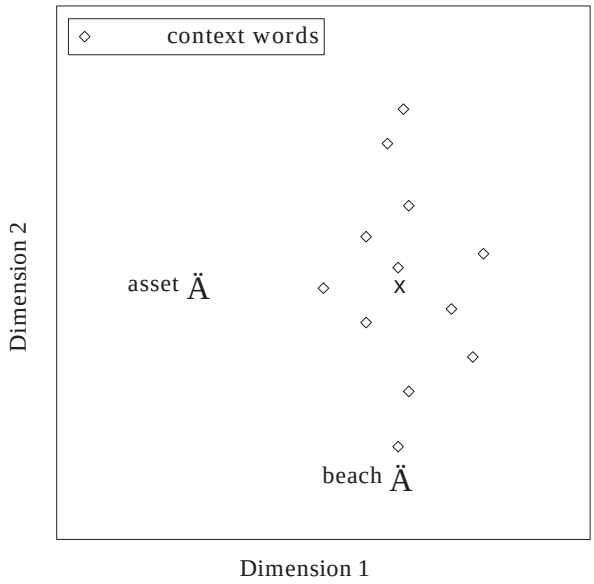

Figure 4-3. Hypothetical context word distribution in a 2-dimensional subspace for example (5a). The centre of the context word 'cluster' is indicated with **x**.

Figure 4-3 displays a hypothetical plot of the context words in (5) in two dimensions of semantic space. Notice that in this contrived example, the two possible completion words *beach* and *asset* are equidistant from the centre of the cluster of context words; an unweighted semantic distance measure would not be able to distinguish these target words in terms of their semantic congruity with the context. But when one considers the dispersion of the positions of the context words along each axis, it is apparent that Dimension 2 is more variable than Dimension 1, and consequently the difference in the location of *beach* from the context word cluster along Dimension 2 should be de-emphasised. If the same general observation holds over all dimensions of semantic space, then this weighting according to contextual relevance would result in *beach* being more 'semantically congruous' with the words in the sentence context than *asset*.

Formally, the Semantic Congruity between a context C and a single target word t is measured as follows. First, the relevance r_i for each dimension i of the semantic space is operationalised as the ratio of the standard deviation s_i of the positions of all context words C along dimension i, over the maximum standard deviation s_{max} for C:

$$r_i = \frac{s_i}{s_{\max}}$$

Next, the vectors representing each word *c* in *C* are temporarily moved to new positions in the space according to a function of *r* and their current distance from *t*:

$$c_i' = c_i + r_i(t_i - c_i)$$

Note that if *r* is large (approaching 1), then any difference between *t* and *C* in the value of component *i* is made less prominent than if *r* is small.[4] Finally, Semantic Congruity (SemCon) is calculated as the mean cosine of the angle between target *t* and each word *c'* in *C*:

$$SemCon(t, C) = \frac{1}{|C|} \sum_{c' \in C} \cos(t, c')$$

The first test of the Contextual Relevance Model was to simulate the basic semantic congruity effect, by capitalising on the fact that highly constrained sentence completion words are also highly semantically congruous with their context. The following two experiments tested whether the distributional characteristics of the words in sentence completion contexts could model the congruity effect: specifically, is this information sufficient to distinguish between words that are highly constrained by the context and frequency-matched controls?

4.2.3 Experiment 6

According to Schwanenflugel and Shoben (1985), high constraint-strength sentences constrain their completions to a small set of words (or perhaps a single word) that meet all the feature restrictions imposed by the context. Processing of unexpected words, which are defined as completions with low cloze probabilities, is not facilitated. For example, for sentence fragment (5) only two completion words had a cloze probability greater than 0.5 (*beach* and *pool*), although 17 different responses were generated in total (Schwanenflugel, 1986). Randomly chosen words should have a low probability of being a member of the small set of words meeting the

[4] *Mahalanobis distance* (Mahalanobis, 1936) is a weighted distance function suitable for multi-dimensional data that takes the variance and covariance of the values in all dimensions into account. Unlike Mahalanobis distance, our method does not consider covariance to be important.

feature restrictions, and thus should be distinguishable from the most expected completion words using the empirical Semantic Congruity measure. The purpose of Experiment 6 was to evaluate the Semantic Congruity measure using a published set of sentence completion norms.

4.2.3.1 Method

The materials for Experiment 6 consisted of 55 high-constraint sentence fragments[5] taken from Griffin and Bock (1998). Cloze probabilities for these sentences had been determined by 25 subjects, using Schwanenflugel's (1986) multiple production norming procedure.

Function words, proper nouns and very low frequency content words (*ie.* having a BNC-spoken lexeme frequency of less than 25) were first filtered from each sentence fragment. Next, control words which matched the 55 constrained completions in BNC-spoken lexeme frequency, syntactic category and length in letters were randomly chosen. The normed completion words had a mean log lexeme frequency of 6.001, compared to a mean of 6.003 for the control words. An example item is given in (6); **bomb** is the highly constrained target word and **agent** is the matched control. (A complete list of the materials is provided in Appendix D.)

(6) The *plane exploded* because of a *hidden* ____. <bomb> <agent>

Finally, high-dimensional vector representations were created for all words (context words, normed completions and controls) from the BNC-spoken using the same methodology and model parameter settings employed in Experiments 1-5.

The Semantic Congruity of each sentence fragment with its constrained completion was estimated as the mean cosine of the angle between vectors, taking into account the variability of the context words along each dimension (*cf.* the procedure detailed in section 4.2.2 above). Dimensions with high variability were weighted less by the measure.

The Semantic Congruity of each sentence fragment with its corresponding control word was similarly estimated.

---

[5] Griffin and Bock (1998) list 60 high-constraint items; five items were eliminated because either their most expected completion word did not meet the frequency threshold of 25 occurrences in the BNC-spoken, or only one context word remained after applying the frequency threshold to all content words in the sentence fragment.

4.2.3.2 Results and Discussion

Using the simple evaluation procedure of scoring an item 'correct' if the normed completion had the higher Semantic Congruity value, the method gave the correct result for 43 out of the 55 sentence fragments. Normed responses were significantly more semantically congruous with their contexts than were the control words: paired $t(54)=4.66$, $p<0.001$.

The results demonstrate that highly-constrained sentence completion words can be distinguished from frequency-matched controls by the corpus-derived Semantic Congruity measure. The distributional properties of the words in the sentence context appear to provide enough information to distinguish the two types of target words in order to successfully model the semantic congruity effect.

4.2.4 Experiment 7

The purpose of Experiment 7 was to further validate the results of Experiment 6, by applying the same procedure to a second set of highly-constraining sentence fragments.

4.2.4.1 Method

55 items were chosen from the topmost part of Schwanenflugel's (1986) sentence completion norms (which were ranked according to the cloze probability of the most expected completion), according to the same frequency criteria employed in Experiment 6.

Co-occurrence vectors for each high-cloze target, randomly-selected matched control word, and the critical context words for each sentence fragment were extracted from the BNC-spoken (see Appendix E for a complete list of the materials). The normed completion words had a mean log lexeme frequency of 5.820 and the controls had a mean frequency of 5.783.

The Semantic Congruity of each context fragment with its normed response and frequency-matched control was estimated as in Experiment 6.

4.2.4.2 Results and Discussion

Experiment 7 closely replicated the results of Experiment 6, using a different set of materials. Using the same evaluation procedure, the Semantic Congruity measure correctly identified 40 out of 55 items. The normed completion words were

significantly more semantically congruous with their contexts than were the control words: paired $t(54)=3.95$, $p<0.001$.

Although not simulations of actual human studies, Experiments 6 and 7 have demonstrated that the semantic congruity effect can be modelled as the relevance-weighted semantic distance between the co-occurrence vector representations of the sentence completion word and the words in the context. The vector for a target word that is highly contextually constrained was 'closer' to the region of semantic space formed by the vectors for the context words, than was the vector for a frequency-matched control.

The results provide further support for the alternative conception of contextual constraint provided by the Contextual Relevance Model. The semantic congruity effect, typically explained in terms of the generation and matching of semantic feature restrictions (Schwanenflugel & LaCount, 1988), can alternatively be described as a function of the similarity of lexical representations created from distributional statistics.

4.3 Feature priming

A small body of related work has been directed at so-called 'feature priming' (*eg*. Moss & Marslen-Wilson, 1993; Tabossi, 1988). In contrast to the research on contextual constraint discussed above, where the effort involved in processing sentence-final words was of interest, feature priming experiments have been aimed at investigating the influence of the sentential context on responses made to a target word presented at the offset of a sentence-final prime word. In these studies, the feature priming context is assumed to make a particular semantic property of the prime salient, which facilitates recognition of a target word possessing that property. The feature priming paradigm thus addresses the influence of contextually relevant aspects of meaning on lexical processing behaviour, which is of interest to this thesis.

4.3.1 Previous research

Research in this area has primarily been aimed at discovering *when* contextual constraints from the semantic context affect recognition of a word. A central question is whether these constraints are able to influence *initial* access to word

meaning: is initial access to semantic information context-dependent or context-independent? Does context-dependence interact with constraint strength, with perhaps only highly-constraining contexts permitting context-dependent effects to emerge (Tabossi, 1988)? Facilitation of a target word containing a semantic property of the prime word that is relevant only in a suitably constraining context would be evidence for context-dependent access of word meaning. Furthermore, if retrieval of semantic information is sensitive to *only* those aspects of prime meaning made salient by the prior context, then processing of a target word that is related to an aspect of prime meaning that is *not* contextually relevant should *not* be facilitated. These issues were addressed in turn by Tabossi (1988) and Moss and Marslen-Wilson (1993).

4.3.1.1 Tabossi (1988)

Using a cross-modal priming paradigm and semantically constraining sentence contexts, Tabossi (1988, Experiment 1) demonstrated support for the context-dependent view of access to semantic information. She proposed that if the preceding linguistic context makes a particular semantic property of a prime word salient, such as [TASTES SOUR] for *lemon* in (7a),[6] then processing of a target word which contains this property (*eg*. **sour**) should be facilitated, compared with presentation of the same target after a neutral sentence context (7b). This hypothesis was verified empirically using lexical decision response time as the behavioural measure of processing effort.

(7) a. The little boy shuddered eating a slice of *lemon*.
b. The little boy was late because he went to buy a *lemon*.

Tabossi's findings are compatible with the Feature Restriction Model of contextual constraint (*cf*. section 4.2.1). Under the assumption that the sentential context imposes semantic feature restrictions on upcoming words, the model predicts facilitation of target words that meet these restrictions. Put another way, if the context builds up expectations that a target word will contain a certain semantic property, then any target containing that property will be facilitated. These semantic expectations would necessarily be dependent on both the context immediately preceding the prime and the prime word itself.

[6] These are Tabossi's translations of her Italian materials.

Tabossi's results are also compatible with the approach to lexico-semantic representation taken in this thesis, and the Contextual Relevance Model of contextual constraint put forward above. If we assume that the aspects of meaning constrained by the context correspond to a region (or subspace) of semantic space where the context words are more tightly clustered (the least variant dimensions of the representational space), then we could associate relevance-weighted Contextual Similarity between prime and target words with response times. The Contextual Relevance Model would therefore predict faster lexical decisions to **sour** presented after (7a) compared with (7b).

4.3.1.2 Moss and Marslen-Wilson (1993)

The context-dependent view of meaning retrieval argued for by Tabossi (1988) is not completely supported by follow-up work by Moss and Marslen-Wilson (1993), who, using a similar cross-modal presentation, demonstrated priming for semantic property targets in both appropriately *and* inappropriately biasing contexts. An appropriately biasing context is assumed to constrain (or restrict) a particular aspect of prime meaning that would not be constrained by an inappropriately biasing context. For example, according to Tabossi, the target word **jungle** should be facilitated when presented after the appropriately biasing sentence (8a) (because the semantic property [LIVES IN JUNGLE] is salient), but not after the inappropriately biasing sentence (8b), compared with presentation after the neutral context (8c). However, Moss and Marslen-Wilson's results indicated facilitation for lexical decision responses to **jungle** after *either* the appropriate or inappropriate context. If the retrieval of word meaning is influenced by semantic property restrictions (or alternatively, by those aspects of meaning made salient by the prior context), then facilitation should not have been found for **jungle** following (8b).

(8) a. The scientist struggled through the undergrowth and tangled vines to take a photograph of the *parrot*.
b. After being cooped up for years, even taking off from the perch was a real effort for the *parrot*.
c. As a birthday present Alison bought her friend a book about a *parrot*.
d. As a birthday present Alison bought her friend a book about a *circus*.

This pattern of results does not seem to be accommodated by either the Feature Restriction Model or the Contextual Relevance Model. Although Moss and Marslen-Wilson's experiments showed that context does have an influence on word

recognition, neither model would predict the facilitation they observed in inappropriately biasing sentence contexts. All that can be concluded from their results is that the context in which a word is processed influences access to its semantic properties. The statistically equivalent priming effect for the appropriate and inappropriate biasing conditions remains unexplained for the moment (we will return to this finding in section 4.3.3).

In summary, the results of the two feature priming studies discussed above indicate that a semantic context that constrains (or makes salient) a particular aspect of meaning (or semantic property) of an unambiguous prime word, can facilitate processing of a target word possessing that aspect of meaning. If the feature priming effect is indeed due to semantic constraints imposed by the context, it should be possible to capture the basic pattern of human facilitation using a 'semantic' measure of similarity derived from distributional information. Chapter 3 demonstrated how the single-word priming effect could be successfully modelled as the semantic distance between high-dimensional lexical representations; by modifying the vector similarity measure to incorporate the distributional properties of the local linguistic context, context-dependent priming effects might also be simulated.

4.3.2 Experiment 8

Experiment 8 was designed to test the Contextual Relevance Model's predictions regarding the influence of feature priming contexts on word recognition. This computational simulation applied the relevance-weighted Contextual Similarity measure (*cf.* section 4.1.3.1) to the experimental materials used by Moss and Marslen-Wilson (1993). The central hypothesis of this reanalysis is as follows: if context succeeds in constraining a semantic property of the sentence-final prime word, then the weighted Contextual Similarity between a target word containing that property and the prime should be greater when the prime is preceded by the constraining sentence context than when preceded by a context which does not constrain that property. Contrary to Moss and Marslen-Wilson's findings, the Contextual Relevance model predicts that facilitation for **jungle** should be observed in the Appropriate condition (8a) only; because the Inappropriate context (8b) does not make the [LIVES IN JUNGLE] aspect of the meaning of *parrot* relevant, the relevance-weighted Contextual Similarity between *parrot* and **jungle** should be equivalent for both the Inappropriate and Neutral (8c) conditions. Finally, in order

to verify that the target and prime words *are* measurably 'close' in the representational space, weighted Contextual Similarity between **jungle** and the unrelated control word *circus* should be the lowest in (8d).

4.3.2.1 Method

A check of the stimulus frequencies in the BNC-spoken indicated that several of the 24 listed items would not be able to be used in the reanalysis, due to their extremely low frequency of occurrence in this corpus. Consequently, we decided to use the entire 100 million word BNC instead, which meant reliable co-occurrence vectors could be created for the entire set of stimuli. Vectors were created from a lemmatised version of this corpus in the usual way, with the same parameter settings used in Experiments 1-7 above.

It is important to ensure that the critical primes and control words were still reasonably matched in BNC frequency, since the cosine metric (on which the relevance-weighted similarity measure is based) is sensitive to vector sparseness, and tends to increase as frequency decreases (*cf.* section 2.3.6). Mean log-transformed lexeme frequency for the prime words was 6.939 compared to 7.246 for the controls, which verified Moss and Marslen-Wilson's original frequency matching procedure.[7]

The single experimental factor, Context Type, consisted of four levels: Appropriate, Inappropriate, Neutral and Control; these conditions are illustrated by (8a-d).[8] Context-dependent semantic similarity between the sentence-final prime and its corresponding target word was estimated for each of the 24 items using the same dimension-weighting procedure described in Experiment 5 above.

---

[7] The slightly higher mean frequency for the control words would tend to bias the estimates of context-dependent similarity for the Control contexts upwards (if it had any effect at all), which is opposite to the experimental hypothesis.

[8] Following Moss and Marslen-Wilson (1993), the Appropriate condition was actually the result of collapsing together, for each item, the relevance-weighted Contextual Similarity for a 'distinctive' property target (*eg.* **jungle**) in its biasing context with the relevance-weighted Contextual Similarity for a 'redundant' target (*eg.* **wings**) in its biasing context. Similarly, the Inappropriate condition combined the data for 'distinctive' targets in the context biasing the 'redundant' property and 'redundant' targets in the 'distinctive'-bias context. Moss and Marslen-Wilson collapsed together the data for the two types of property targets in this way after finding that their distinction failed to interact with any other experimental factor.

Table 4-2. Weighted Contextual Similarity Between Prime and Target by Type of Context.

Context Type	Contextual Similarity Mean	SD
Appropriate	0.439	0.224
Inappropriate	0.382	0.168
Neutral	0.376	0.174
Control	0.308	0.181

4.3.2.2 Results and Discussion

The mean relevance-weighted Contextual Similarity measurements for each condition are given in Table 4-2. As anticipated, the highest prime-target similarity was found for the Appropriate condition, and the lowest for the Control condition. The effect of the Context Type factor was significant: $F(3,69)=6.58$, $p<0.001$. Planned multiple comparisons (with alpha levels adjusted to compensate) indicated that weighted Contextual Similarity for the Appropriate contexts was significantly higher than for the Neutral contexts: $F(1,23)=7.60$, $p<0.05$. There was no reliable difference between the Inappropriate and Neutral contexts, $F(1,23)<1$, and the difference in similarity between the Appropriate and Inappropriate contexts did not reach significance: $F(1,23)=3.95$, $p>0.15$.

The results of this computational reanalysis of Moss and Marslen-Wilson's materials are in accordance with the predictions of the Contextual Relevance Model: facilitation was found for target words in Appropriate contexts only. Generally speaking, the dimension-weighting procedure succeeded in determining those aspects of meaning relevant to the words in the sentence context, which allowed context-dependent semantic relatedness to be easily estimated. However, the simulation has also demonstrated that the Contextual Relevance Model makes incorrect predictions about human performance, since *contra* the model, Moss and Marslen-Wilson also observed facilitation for property targets presented after an Inappropriately biasing context.

4.3.3 Is feature priming actually contextual priming?

A plausible interpretation of the results of both Tabossi (1988) and Moss and Marslen-Wilson (1993) is that priming effects were the result of contextual

facilitation from not just the designated sentence-final prime, but from the sentential context as a whole. It might be the case that a sufficiently constraining context could *directly* facilitate a target word related to the meaning invoked by the words in the context, without the need for a related prime word at all (*eg.* Williams, 1988). In this view, the cross-modal priming observed by Tabossi (1988, Experiment 1) is a 'global' effect – facilitation is due to the semantic properties of the context as a whole – rather than a 'local' effect from the sentence-final noun. However, further experimental results (Tabossi, 1988, Experiment 2) apparently rule out this contextual priming explanation. No priming effect was found for property targets when the sentence context was terminated by a syntactically compatible, but semantically unrelated prime word. For example, facilitation to the target **fat** was found when presented after (9a), but not after (9b).

(9) a. To follow her diet, the woman eliminated the use of *butter*.
b. To follow her diet, the woman eliminated the use of *wine*.

Although Tabossi argues that the presence of the prime word *butter* is necessary to demonstrate the context-dependent pattern of priming, the possibility remains that an unrelated prime might disrupt (in some sense) the semantic constraints imposed by the context up to that point. That is, *wine*, but not *butter*, might impede (or block) any contextual priming effect. It would have been interesting to compare response facilitation of the property target **fat** presented after the sentence-final prime *butter*, to processing of the same target when presented at the offset of the penultimate word in the sentence. Regardless of the results of such an investigation, contextual facilitation appears to be eliminated as an explanation of context-dependent priming for Tabossi's materials.

The hypothesis that feature priming can be attributed to priming from the sentential context preceding the prime word was further discounted by Moss and Marslen-Wilson (1993, Experiment 1). Recall that this study revealed statistically equivalent facilitation for semantic property targets in both appropriate and inappropriately biasing contexts (*eg.* **jungle** presented after either [8a] or [8b]). Moss and Marslen-Wilson argue that "… priming in the inappropriate condition could not result from direct contextual facilitation as there was very little relation between the context sentence and the target word" (p. 1269). This claim is based on the authors' care to exclude strong normative associates of the target word from the

sentence context, and additionally by saliency ratings elicited in a material selection pretest.

Note that the argument against contextual priming is based on the one hand by the *lack* of observed facilitation when the feature priming context was appropriate for the property target, but the prime word was unrelated (Tabossi, 1988, Experiment 2), and on the other by the *presence* of a priming effect when the constraining context was supposedly inappropriate for the property target and the prime was related (Moss & Marslen-Wilson, 1993). In order to reconcile this disparate evidence for the same argument, we should consider the differences between the contexts used in the two experiments. First of all, they might differ in constraint strength: the degree that those aspects of meaning relevant to the property target are constrained. Tabossi's sentences were quite short and perhaps would require the support of the related prime word in order for a contextual priming effect to emerge. Although Moss and Marslen-Wilson's sentence contexts were longer and therefore could conceivably constrain those aspects of meaning possessed by the Appropriate property target, even without the designated prime word being present, this would not explain the facilitation obtained in the Inappropriate condition. The problem might lie with their materials pre-test. It is not at all clear that off-line saliency ratings reflect the critical variables responsible for contextual priming effects. As long as the Inappropriate sentence context (8b) up to the penultimate word succeeded in constraining the property [LIVES IN JUNGLE] to a greater extent than the Neutral context (8c), then this counterintuitive pattern of priming would be accounted for.

It is possible that this relationship between Inappropriate context and property target might be revealed using the corpus-derived Semantic Congruity measure developed in section 4.2.2 above. For example, the target word **jungle** might be as semantically congruous with the Inappropriately biasing context as with the Appropriately biasing context, and less congruous with the Neutral context. If so, an alternative explanation for Moss and Marslen-Wilson's results could be advanced: a semantic property target is facilitated as long as it is semantically congruous with the sentential context, irrespective of the rated salience of the property in that context. Feature priming would therefore be explained as facilitation that is attributable to the sentential context as a whole.

Although the results of such a test would only be suggestive of semantic congruity as an *explanation* (because there is no independent measure of semantic congruity

available against which the corpus-derived measure could be validated), we nevertheless tested this hypothesis using the same computational procedure employed in Experiments 6 and 7 above, with Moss and Marslen-Wilson's (1993) materials.

Considering the content words in the context up to, but not including the sentence-final prime, the mean Semantic Congruity of a target word with its Appropriate context was 0.252 (SD=0.104), and with its Inappropriate context was 0.227 (SD=0.105). In comparison, the mean Semantic Congruity of a target with its Neutral context was 0.179 (SD=0.048). An ANOVA revealed a significant effect of the Context Type factor: $F(2,46)=6.03$, $p<0.01$. Planned comparisons indicated a reliable difference between the Appropriate and Neutral conditions, $F(1,23)=16.81$, $p<0.01$, and a marginally significant difference between the Inappropriate and Neutral contexts, $F(1,23)=5.04$, $p<0.105$. The difference in Semantic Congruity between the Appropriate and Inappropriate conditions was not significant: $F(1,23)=1.05$.

In summary, the results of this computational reanalysis approximate the observed pattern of human performance, and is consistent with the hypothesis that contextual facilitation between the property target and individual words in the context was responsible for the unanticipated pattern of priming reported by Moss and Marslen-Wilson. Compared with the Neutral context, Moss and Marslen-Wilson's target words were 'closer' in corpus-based representational space to both their Appropriate and Inappropriate contexts. Priming of semantic property targets in the feature priming paradigm can be interpreted as a 'global' effect of contextual facilitation from individual words in the preceding context, as opposed to a 'local' effect from the final word in the sentence.

Experiment 8 has illuminated a deficiency in the initial modelling assumption made here, namely that feature priming could be simulated using a measure of the representational similarity between two *words*. This discrepancy between model predictions and the empirical data is apparently resolved if one accepts that words in the context preceding the designated sentence-final prime are also able to influence the processing of the semantic property target. Even though Moss and Marslen-Wilson's materials were carefully controlled to avoid associative relationships between context words and targets, the fact that their property targets were equivalent in Semantic Congruity with both the Appropriate and Inappropriate contexts invites reinterpretation of their results. It now seems entirely

plausible that the observed facilitation of a semantic property target was not simply due to its relationship to the designated prime word, but was also dependent on the context preceding the prime. Moss and Marslen-Wilson's pattern of results is compatible with a contextual priming explanation, indicating that future feature priming experiments need to control for the influence of the distributional information carried by the words in a sentence context.

4.4 Limitations of the Contextual Relevance Model

Experiments 6-8 have shown that the corpus-based Contextual Relevance Model possesses several desirable properties. First, it is *objective*: it makes similar behavioural predictions to feature-based theories such as the Feature Restriction Model, without stipulating an inventory of semantic features. Second, it is psychologically plausible; this is demonstrated by its ability to capture both the semantic congruity and feature priming effects. Third, it is compatible with the representational approach to modelling the single-word priming effect investigated in Chapter 3. Despite these positive features, the model makes some unrealistic assumptions.

One potentially serious deficiency of the Contextual Relevance Model of contextual constraint is that a context is treated as an unordered set – a 'bag' of words, to use the computational linguistics terminology. The relationship of the target word to the lexical items in the context in terms of *temporal proximity* is not considered by the model; additionally there is no role attributed to the *order* of the context words. In other words, the Contextual Relevance Model ignores possible high-level effects on processing behaviour from syntactic structure (*ie*. syntactic dependencies holding between context words), as well as effects due to computational resource constraints (such as the capacity of working memory; see Caplan and Waters, 1999). The reason why the model was developed using the 'bag-of-words' method was due to computational simplicity and the fact that this sort of approach has been successful both in Information Retrieval and for the task of word sense disambiguation in NLP (*eg*. Schütze, 1998).

Order may also be a critical parameter, meaning that words in the context should not be considered to have influences independent of their relative ordering. For instance, it is possible that a context word could influence a subsequent word's contribution to the semantic congruity of a sentence fragment with a completion

word. The importance of syntactic order for human processing behaviour has been established by studies comparing word recognition performance in scrambled and unscrambled contexts (*eg*. Masson, 1986; Vu, Kellas & Paul, 1998). Contrary to these findings, the Contextual Relevance Model's 'bag-of-words' representation of context entails predictions of equivalent response time facilitation for *beach* presented in contexts (10a) and (10b), which clearly would not be observed empirically.

(10) a. On a hot summer's day many people go to the *beach*.
b. Many to summer's on go hot the a people day *beach*.

Gallant (1991) proposes a simple procedure for taking relative word position within the context into account. His idea is to weight the contribution of each lexical item in the context according to its distance (in word units) from the target word, with context words immediately preceding the target assigned the most weight. This method would appear to reflect (albeit very crudely) the computational limitations of working memory, but of course it is still insensitive to the presence of syntactic relationships between context words.

The relevance of parameters such as temporal proximity and order to the representation of context, particularly for measures of distributional similarity, needs to be addressed in future research. Nevertheless, it is certainly interesting that the simple 'bag-of-words' approach succeeds in capturing a substantial amount of human behavioural data. In Chapter 6, we develop a model of *incremental* processing which is sensitive to temporal proximity and order.

4.5 Summary

In this chapter, we have stretched the limits of our representational model of word meaning in order to account for the impact of the local linguistic context on both interpretation and lexical processing behaviour. We provided evidence for the context-dependent nature of word meaning from work in lexical semantics, which motivated the development of a computational procedure for modifying the Contextual Similarity measure in order to take the distributional characteristics of the local context into account. This approach proved successful: Experiment 5 demonstrated that intuitions about the contextual appropriateness of a particular meaning of an ambiguous word could be captured using the measure. Next, we

reviewed the literature on contextual constraint, and in Experiments 6 and 7, demonstrated that the basic semantic congruity effect could be captured using the relevance-weighted version of the Contextual Similarity measure. The phenomenon of feature priming seemed an obvious candidate for further testing of the approach; the results of Experiment 8 indicated that the behavioural data could be modelled successfully by replacing the assumption that feature priming is a 'local' effect (from the sentence-final word), with the view that facilitation is dependent on the context preceding the prime. This explanation accounted nicely for the counterintuitive findings of Moss and Marslen-Wilson (1993). Finally, we identified some inherent limitations of the representational approach to modelling context-dependent processing behaviour.

5.
Lexical Processing in the Absence of Context

The psychological relevance of co-occurrence statistics to language processing phenomena has been established in Chapters 3 and 4. In this chapter, we take this investigation further, by proposing a role for how distributional information relates to lexical processing behaviour in the *absence* of context, such as in the task of recognising an isolated word. We describe a new lexical property, contextual distinctiveness (CD), and formally define it by imposing an information-theoretic interpretation onto the co-occurrence data making up a word's high-dimensional vector representation. Initial sections deal with issues concerning the measurement properties of CD: setting the parameters of the underlying representational model, and the reliability of the measure. The remainder of the chapter is devoted to empirical investigation: evaluating the psychological plausibility of CD as a predictor of lexical processing effort, and closely examining the relationship between CD and other lexical properties such as concreteness and ambiguity.

5.1 An information-theoretic measure of contextual behaviour

In the model of lexico-semantic representation presented in the previous chapters, the relationship *between* high-dimensional lexical representations (*ie*. vector similarity) was the quantity of interest. What is the nature of the information contained in a *single* word vector? In this section, we investigate how words can be distinguished according to their distributional properties.

A co-occurrence vector summarises a word's distributional profile – which words it tends to occur with, and how often. For instance, if word *w* tends to appear in a

wide variety of linguistic contexts (*eg. run*), we would expect the distribution of the words it occurs with to be rather diffuse. Conversely, if *w* typically appears in a small number of different contexts (*eg. amok*), or is perhaps found in diverse contexts but is much more common in a subset of them, its contextual distribution would be less diffuse and we could describe *w* as *contextually distinctive*. Seen from another perspective, encountering the word *run* in isolation is not particularly informative about the linguistic contexts it occurs in (since *run* can appear in wide range of contexts), whereas observing *amok* almost certainly brings to mind the verbal context *run*. Because words appear to vary according to informativeness about their contexts of use, this property, contextual distinctiveness, can be treated as a continuous variable.

We use the relative entropy measure from information theory to quantify the subjective notion of contextual distinctiveness (CD). Since CD intuitively refers to the amount of *information* provided by word *w* about its contexts of use, it is operationalised as the relative entropy (or Kullback-Leibler distance) between the distribution of context words occurring in a window of words around *w* (the *posterior* distribution), and the distribution of context words expected when *w* is not taken into account (the *prior* distribution). CD can be understood as the quantity of information conveyed about *w*'s contextual behaviour.[1]

Note that the posterior distribution is simply *w*'s co-occurrence vector representation, with counts[2] converted to conditional probabilities, and the prior is the distribution of context words based on their estimated independent probabilities of occurrence (or relative frequencies) in a large corpus. The prior can be interpreted as the probability distribution expected if the corpus had been created from word tokens randomly chosen according to their relative frequencies in a real corpus. Such a corpus (unlike natural language) would be completely unstructured, and consequently co-occurrence vectors extracted from this corpus would simply record the relative frequencies of the context words (*ie.* no linguistic dependencies between words would be encoded).

---

[1] It is tempting to state that CD measures the amount of *semantic* information conveyed by a word, but as Chater (1989) argues, information-theoretic tools can only provide quantitative, not qualitative answers. They do not permit anything to be said about the *content* of the information being conveyed.

[2] Encoding co-occurrence using counts instead of the log-likelihood ratio statistic allows a vector to be easily converted to a conditional probability distribution and, relevant to the Bayesian update procedure presented in Chapter 6, greatly simplifies computation by allowing a word vector to be interpreted as a multinomial likelihood function.

The formal definition of CD is as follows. First, for the prior distribution, we define C to be a discrete random variable ranging over the alphabet of symbols $\{c_1, \ldots, c_n\}$, with probability mass function $P(c)$. (The alphabet is the finite set of context words that label the components of a co-occurrence vector.) The values of $P(c_i)$ are obtained using the maximum likelihood estimator:

$$P(c_i) = \frac{f(c_i)}{\sum_{j=1}^{n} f(c_j)}$$

In this equation, $f(c_i)$ is the frequency of c_i in the corpus, and the denominator is the summed corpus frequency of the n words in the alphabet. Note that the prior distribution is a true probability mass function, since $P(c_1)+P(c_2)\ldots+P(c_n)=1$.

We next describe the posterior distribution (the distribution of context words given that target word w occurs) using the probability mass function $P(c|w)$. Each conditional probability $P(c_i|w)$ is derived from the co-occurrence frequency of the target word w with the context word c_i, normalised by the total co-occurrences of w with each possible symbol c:

$$P(c_i|w) = \frac{f(c_i, w)}{\sum_{j=1}^{n} f(c_j, w)}$$

Finally, CD is calculated as the relative entropy between the two probability mass functions $P(c)$ and $P(c|w)$, using the convention $0 \log 0 = 0$ (justified by continuity):

$$D\big(P(c) \,||\, P(c|w)\big) = \sum_{i=1}^{n} P(c_i|w) \log_2 \frac{P(c_i|w)}{P(c_i)}$$

Formally, CD measures the quantity of information provided about a random variable (the contexts that word w appears in) by an event (observing word w). Since the above equation uses base 2 logarithms, the units of information are expressed in *bits*.[3]

[3] It is important to note that the CD measure is implicitly conditioned on the parameters used when constructing co-occurrence vector representations: window size, the selection of context words, and the choice of corpus. Varying these parameter settings will yield different values of CD for the same word, to a limited extent.

The present formulation of CD using relative entropy is highly similar to Resnik's (1993) definition of the selectional preference strength of a predicate for its arguments. However, Resnik defines the prior and posterior distributions in terms of the taxonomically-organised semantic classes in WordNet (Miller, Beckwith, Fellbaum, Gross & Miller, 1990), rather than using a finite set of co-occurring words, as is done here. His central motivation for using classes instead of words is so that the selectional association of a *particular* predicate-argument pair can be estimated, for the cases where there is unreliable (or no) corpus evidence. For purposes of psycholinguistic modelling, we consider lexical, as opposed to class, co-occurrence probabilities to be the distributions of interest. Although Resnik's method seems promising for tasks in natural language processing such as word sense disambiguation (Resnik, 1997), its main drawback is that the estimation of class probabilities assumes *a priori* semantic categories, whereas collecting lexical co-occurrence statistics does not presuppose the existence of a cognitive semantic organisation system analogous to the WordNet taxonomy.

5.1.1 CD and lexical processing effort

What relation to human language processing would we expect CD to have? Adopting the uncontroversial view that the primary function of human language comprehension is the efficient recovery of meaning from an utterance, it seems reasonable that the processor is adapted to minimize the cost of retrieving semantic information from memory, by forming expectations about the meaning of upcoming words (*cf.* section 1.1.1). Under this assumption, the cost of recovering the meaning of word *w* would vary directly with the *difference* between the processor's predictions about the meaning of *w* and the 'actual' meaning of *w*. Put in a slightly different way, lexical processing difficulty is assumed to correspond to the difference between what the processor expects to find and what it actually finds.

Since the general approach to semantic representation taken in this thesis equates the meaning of a word with its contexts of use, recovering a word's meaning amounts to retrieving its associated distributional information. In the case of recovering the meaning of an isolated word,[4] we assume that the lexical processor starts off in a state where it has minimal expectations about its distributional characteristics (*ie.* its meaning). The best it can do is anticipate a distributional

[4] Or, presumably, recovering the meaning of the first word in a discourse. The two situations are identical from the point of view of the model.

profile that is in some sense neutral – a distribution based on the relative frequencies of the context words.

Upon encountering the target word, the processor's expectations are compared to (or, alternatively, are updated with) the target's distributional representation. Changing from an uninformative, neutral prior state to a state that is minimally different conveys only a small amount of information, involves minimal modification to initial predictions, and incurs little processing cost. Moving to a very different state conveys much more information, involves substantial amendments to the processor's original hypothesis, and incurs a higher processing cost. The CD measure provides an quantitative estimate of this processing cost; we propose that CD – the amount of information conveyed by *w* about its contexts of use – is predictive of the effort involved in recovering the meaning of *w*, in the absence of context. CD is defined in section 5.1 above, and we can measure lexical processing effort using any number of laboratory techniques, such as recording the time required to recognise a word presented in isolation. This hypothesis can now be tested.

If the CD measure successfully captures a substantial portion of the between-word variability in the effort of recovering word meaning, then the basic behavioural prediction is that words with a high CD value should be more difficult to process than words with a low CD value. Experimental tasks that have been claimed to involve the access and instantiation of word meaning, such as lexical decision, semantic categorisation, pronunciation, and translation could supply a testing ground for this hypothesis. If our claims are correct, we would expect to find a relationship between CD and a behavioural measure of processing effort such as lexical decision response time. This prediction is tested in Experiments 9 and 10 below.

As discussed in Chapter 2, creating lexical representations from co-occurrence statistics requires decisions to be made about a number of model parameters. The CD measure requires estimation of two probability distributions; CD values will vary depending on how model parameters are set. As a result, the success of CD as a predictor of lexical processing effort would also vary. We set the parameters of the representational model empirically, by recording the amount of visual lexical decision (VLD) response time variance explained by CD, for a sample of words. The window size and the number of context words used to define the representational space are two of the most important parameters (*eg*. Levy, Bullinaria & Patel, 1997); consequently the parameter exploration procedure

described below examined the effect of a range of values for each variable on the prediction of lexical decision latencies.

5.1.2 Parameter optimisation

Response time data from a recent VLD experiment[5] were used to evaluate the effect of varying model parameters. From the list of 165 words used in this study, the 70 content words with the smallest response time variance were selected for the optimisation procedure (see Appendix F). Next, CD was calculated for each word using the lemmatised version of the BNC-spoken, and a simple linear regression was conducted between CD and VLD latency. The coefficient of determination (r^2) served as the measure of goodness-of-fit. The amount of sample response variance accounted for by CD was then recorded for each location in the parameter space.

We examined seven window sizes ranging from ±1 to ±10 words, and varied vector size from 250 to 3000 context words, counting from the top of the lexeme frequency list containing content words only. Forty-nine parameter combinations were investigated (see Figure 5-1). From the plot it is apparent that r^2 is not a monotonic function of either variable; for example, the middle range of the vector size parameter is worse than the high or low values. The best fit to the response time data was achieved using 500 context words to define the semantic space, with a window size of around ±5.[6] These settings were verified by Friedman non-parametric analyses of variance carried out separately for each variable. Window size had a significant impact on goodness-of-fit: F_r=30.72, 6 *df*, p<0.0001, with the largest mean rank obtained for the window size of ±5 words. However, multiple comparisons (Siegel & Castellan, 1988) indicated this window size to be reliably better than the ±1 and ±2 word windows only, at the α=0.05 level of significance. The number of context words used to calculate CD also had a significant effect: F_r=34.16, 6 *df*, p<0.0001, with the largest mean rank for 500 context words. The multiple comparison test indicated that this dimensionality gave significantly better performance than spaces constructed using 1000 and 1500 dimensions only.

[5] Data were collected as part of an ESRC-funded research project, in collaboration with Richard Shillcock, Peter Hipwell and Will Lowe of the Centre for Cognitive Science. Peter Hipwell ran the experiment and performed the data trimming. See section 5.2.1 for a complete description of the method.

[6] Note that these 'optimal' settings were not dependent on this particular set of 70 words. Parameter spaces plotted for three sets of 70 items randomly chosen from the 123 content words included in the VLD experiment proved to be very similar to the parameter space derived using the current 70 items.

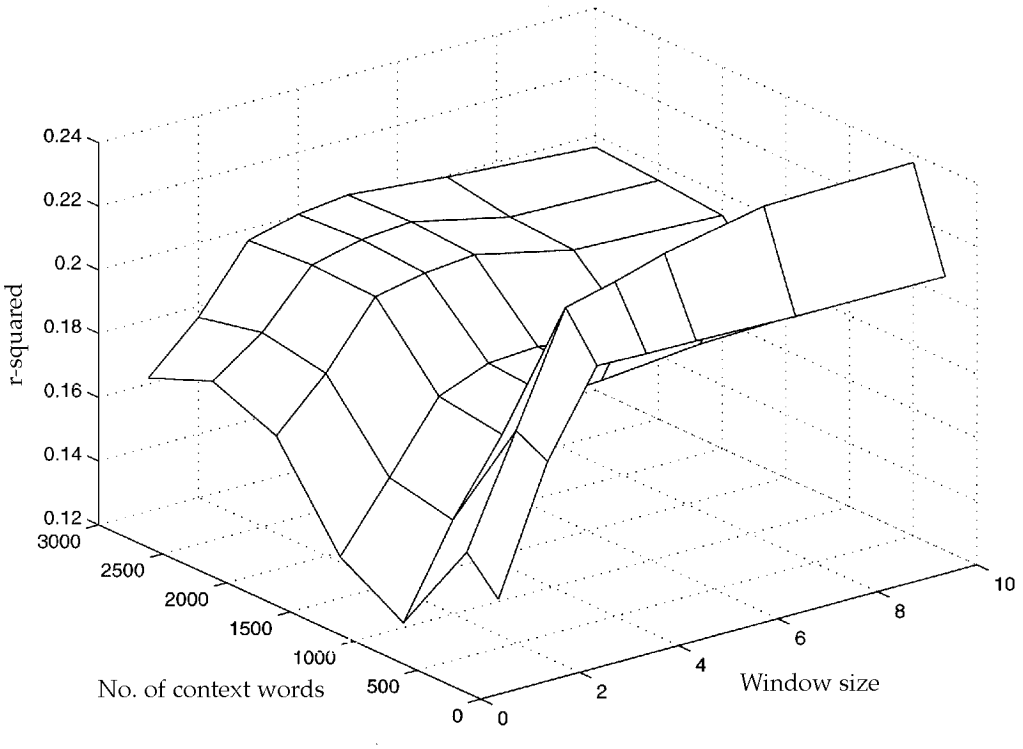

Figure 5-1. Response time variance accounted for by CD calculated using different parameter settings.

These parameter settings (±5 word window, 500 most frequent content words serving as context words) were employed in all subsequent modelling. Because co-occurrence is encoded using counts, and settings were optimised using response data, these settings are slightly different to those used in Chapters 3 and 4.

Now that an 'optimal' set of parameter settings has been determined, it is possible to calculate CD for any word in the corpus. However, it would be useful to verify that CD really does capture the subjective concept of contextual distinctiveness described earlier. The spoken language component of the BNC contains disfluencies such as filled pauses (*ah*, *erm*, *hmm*, etc) – these would be expected to be among the least contextually distinctive tokens in the corpus, since they can occur in virtually any context. This was verified in terms of the objective CD measure; *er* and *erm* received the first and third lowest CD scores of the entire lexicon (0.041 and 0.046 bits, respectively).

In order to graphically illustrate the distributional differences that underlie different CD values for words matched on other lexical properties, we selected two words of equivalent lexeme frequency and plotted their prior and posterior probability distributions (Figure 5-2 and Figure 5-3). *Lane* and *customer* are both

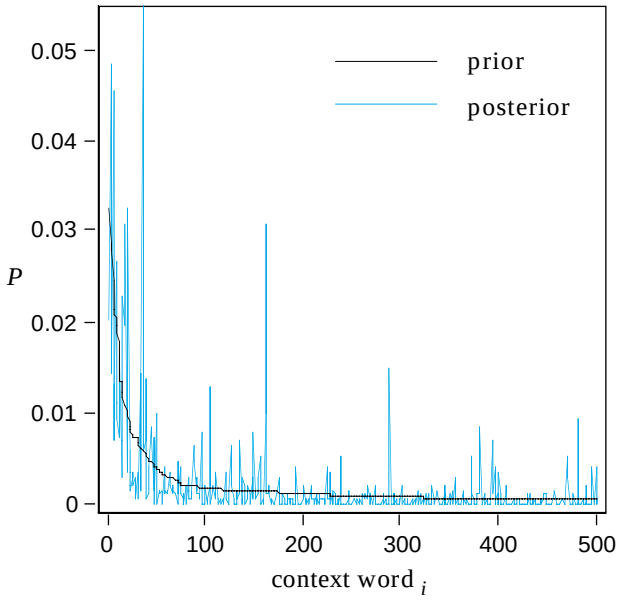

Figure 5-2. Prior and posterior probability distributions for *lane*. (For clarity, lines are used instead of bars.)

unambiguously nouns, according to CELEX, and have BNC-spoken lexeme frequencies of 613 and 614, respectively. However, *lane* is substantially higher in CD than *customer* (1.027 bits *vs*. 0.524 bits). Note that the posterior distribution of *lane* diverges from the prior to a greater degree than does the posterior of *customer*, reflecting, in part, the fact that *lane* occurs in a number of common collocations, such as *back lane* and *fast lane*.

5.1.3 Reliability of the CD measure

A co-occurrence vector records the distributional profile of a word in a *corpus*, and is therefore merely an estimate of its 'true' distribution in unlimited natural language. Since calculation of CD depends crucially on the distributional information contained in a sample of language, it is necessary to assess its *reliability*. Given access to a second corpus of similar size, how comparable are CD values calculated using each corpus?

CD values computed for common words will be more reliable than for rare words for the same reason that corpus frequency is more reliable for common words: the larger the sample, the more accurately the population value can be estimated (the smaller the measurement error). Two separate estimates of CD for the same high frequency word will be close to the population value, and will therefore be very

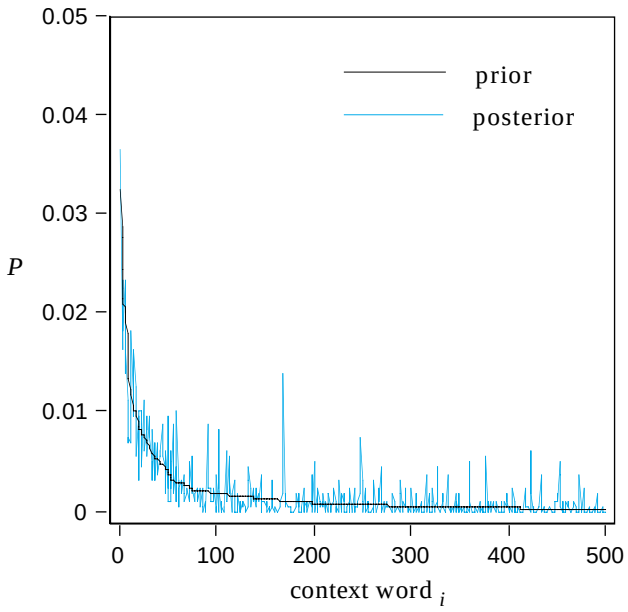

Figure 5-3. Prior and posterior probability distributions for *customer*.

similar. Because of the unreliability of statistics based on small samples, the results of psycholinguistic studies that use rare words as stimuli are suspect, particularly those using sets of words matched on corpus frequency (Lovelace, 1988). For instance, Gernsbacher (1984) demonstrated that low-frequency words varied substantially in experiential (subjective) familiarity, leading her to reinterpret the results of several word recognition experiments.

It would be useful to determine a practical lower limit on the frequency of words for which CD can be confidently measured. Reliability of the CD measure could be estimated by comparing the current CD values obtained using the BNC-spoken to values computed from another corpus of spoken language, for the same set of words.[7] The correspondence between the measurements should be strongest for high-frequency words and fade to nonsignificant levels as frequency drops.

Lacking another comparable speech corpus, we decided to split the BNC-spoken in two and calculate CD for samples of words taken from a range of frequency intervals. This was the strategy followed in Chapter 2 to estimate the reliability of co-occurrence vectors.

[7] Comparing the word vectors themselves using Kendall's *W*, for example (*cf.* section 2.4.3), constitutes a stricter test, but because CD is formulated as the 'distance' between distributions, it is theoretically possible for two quite differently shaped posterior distributions to yield identical CD scores.

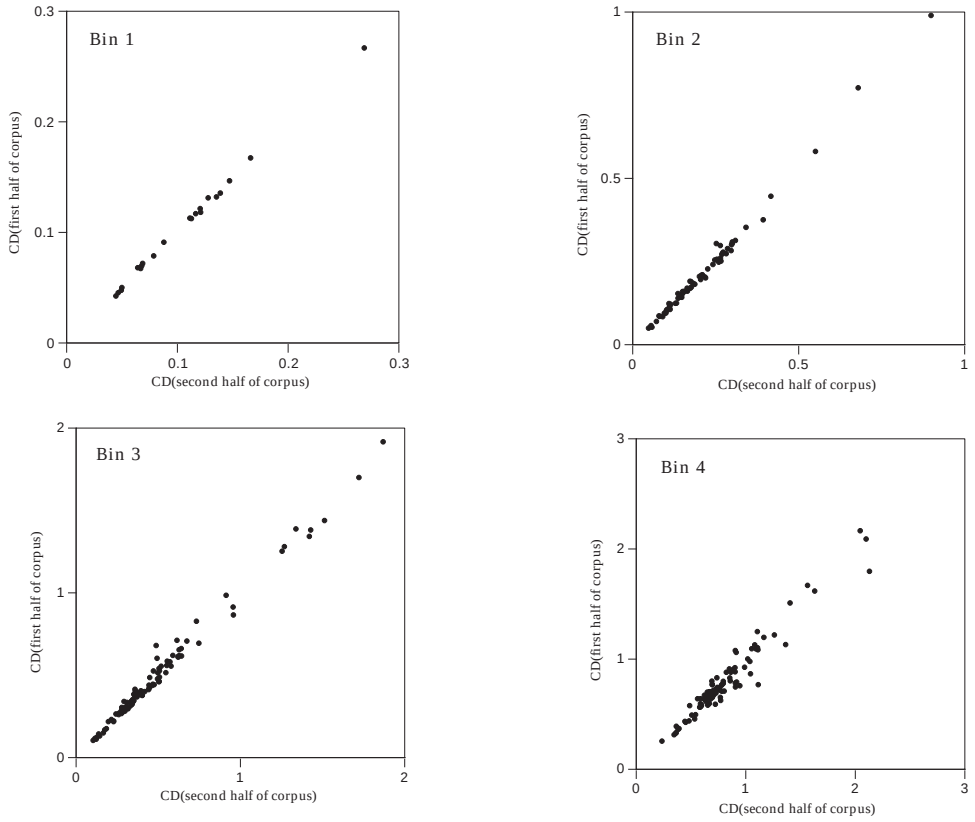

Figure 5-4. Scatter plots of CD calculated from two subcorpora for the topmost four frequency intervals.

Adopting the same eight frequency intervals and random samples used in section 2.4.3, we calculated CD for each word (690 in total) from each half of the corpus separately, and plotted the association between CD scores. It is clear from Figure 5-4 and Figure 5-5 that as word frequency decreases, the linear relationship between the two sets of CD values deteriorates.

Kendall's coefficient of concordance *W* was used to estimate the reliability of CD for each sample. Recall that *W* will be high when the rank orders of the CD scores derived from each subcorpus are highly similar. As expected, reliability was very high for the bins containing the most frequent words (see Table 5-1). The significance of the reliability score for each bin was tested with the X^2 statistic. At the α=0.01 level of significance, the null hypothesis that the reliability scores for bins 7 and 8 (the two lowest frequency intervals) were due to chance failed to be rejected.

What did this exercise tell us about the reliability of CD? Clearly, the measure is the most reliable for words for which the most corpus evidence exists. Paralleling

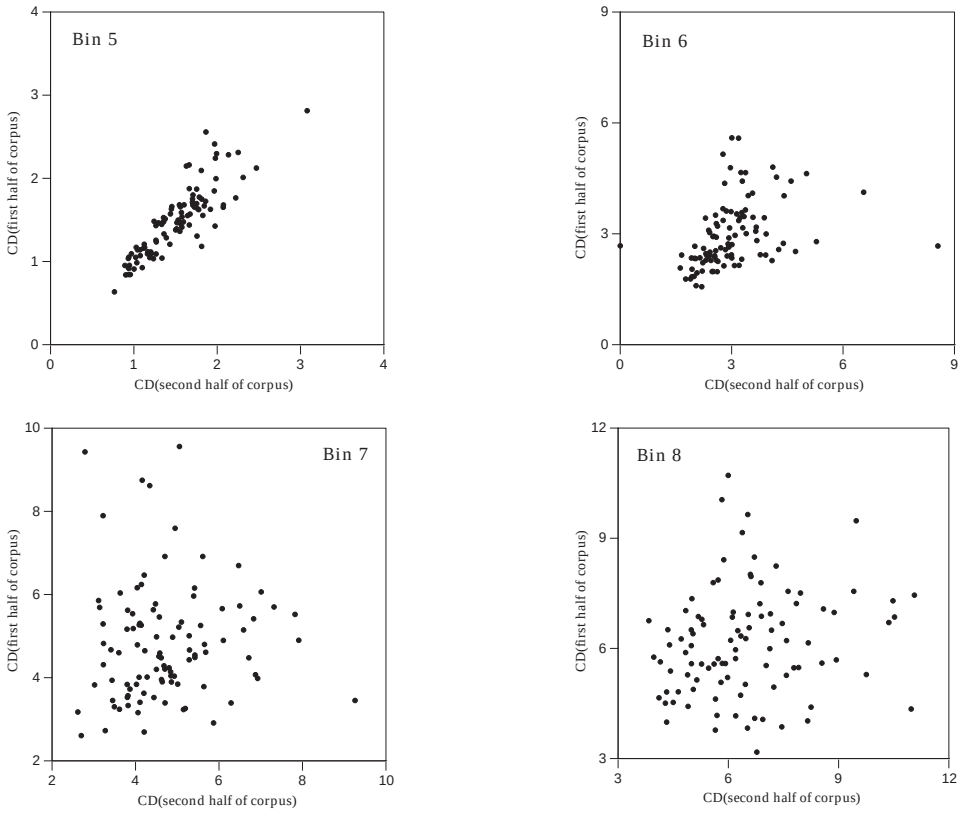

Figure 5-5. Scatter plots of CD calculated from two subcorpora for the lowest four frequency intervals.

the results of Chapter 2's vector reliability investigation, CD also could not be reliably measured for words falling in the two lowest bins. Consequently, CD was not calculated for words with a lexeme frequency of less than 25 occurrences, restricting its applicability to approximately 8,000 lexemes in the BNC-spoken.

5.2 CD and word recognition

5.2.1 Experiment 9

As a first test of the proposed role of CD as a predictor of lexical processing effort, we examined the ability of CD to account for human performance on a visual word recognition task. Recognising the individual words in a sentence is a necessary first step for the high-level comprehension processes concerned with reading; the process

Table 5-1. Reliability of CD for Eight Samples.

Bin	Log Frequency Range	1st Word in Bin	N	Sample Size	Kendall *W*	X^2
1	12.963-11.344	be	21	21	0.998	39.92*
2	11.343-9.724	yeah	69	69	0.996	135.42*
3	9.723-8.103	work	195	100	0.994	196.73*
4	8.102-6.483	case	810	100	0.971	192.20*
5	6.482-4.862	goal	2204	100	0.939	185.86*
6	4.861-3.242	valid	4624	100	0.789	156.21*
7	3.241-1.621	zebra	9150	100	0.566	112.07
8	1.620-0.000	zulu	28378	100	0.604	119.51

* Significant at α=0.01

of word recognition arguably involves retrieving the semantic information associated with these words. If CD succeeds in capturing a non-trivial portion of the processing effort involved in recovering word meaning, then the prediction is that the time taken to identify a string of letters as a valid word should vary directly with the quantity of information conveyed by that word about its contexts of use. In order to address this hypothesis, Experiment 9 employed the standard visual lexical decision (VLD) task to produce a dataset of response times for a large selection of words. We analysed this dataset using linear regression techniques, treating response time (RT) as the dependent variable. Besides CD, two other variables generally considered to be good predictors of VLD latencies were included in the analyses: word length in letters (WL) and word frequency. Frequency was measured using logarithmically-transformed lexeme frequency (lnLF) from the BNC-spoken.

5.2.1.1 Method

Eighteen subjects made timed word/nonword decisions to a total of 165 visually-presented word stimuli, together with an equal number of nonwords. The word stimuli were drawn from a variety of syntactic categories and ranged in length from four to seven letters. BNC-spoken lexeme frequency ranged from 6 to 21,723 occurrences per million. Nonwords were constructed by altering three letters of each of the 165 word stimuli, such that each nonword was one letter away from a real English word.

PsyScope experimental software (Cohen, MacWhinney, Flatt & Provost, 1993) was used to present the stimuli and record lexical decision responses. Each trial consisted of a centrally-presented fixation point displayed for 500 ms, followed by the experimental item. Immediately after a response was made the next trial was

initiated. The 330 stimuli were presented in three blocks, with a rest break in-between each block. Order of presentation was randomised separately for each participant.

5.2.1.2 Results

Errors were replaced with the subject's mean RT; there were no errors due to response timeout (set at 2000 ms). Response latencies longer than two standard deviations (SDs) above a subject's mean RT were trimmed to two SDs. Out of a total of 2,970 responses, 104 lexical decision errors were replaced, and 129 RTs were trimmed (3.5% and 4.3% of the data, respectively).

Of the original 165 items, about one-quarter were function words (according to their canonical CELEX classification; *cf.* section 2.3.3); we did not include these words in the analyses. Although a CD value can be calculated for any word, we did not attempt to predict response times to function words for two reasons. First, compared with content words, function words convey little meaning (by definition, function words are primarily concerned with grammatical function and syntactic structure [see Cann, *in press*]), and a measure that purports to model the process of building expectations about meaning would not be expected to have any predictive power for this class. Second, in English at least, reaction time data for function words and other very high-frequency words is subject to a floor effect (Gordon & Caramazza, 1985), which means that there is insufficient variance in this range to allow a meaningful regression analysis.

After removing the 70 items used for parameter optimisation (*cf.* section 5.1.2), all function words, and two inflected forms from the dataset, 53 items remained. The critical materials are listed in Appendix G. The descriptive statistics for RT and the three predictor variables are given in Table 5-2, and Table 5-3 displays the results of a correlation analysis.

Table 5-2. Descriptive Statistics for RT and Three Predictor Variables.

	RT	WL	lnLF	CD
Min	465	4	4.127	0.091
Max	576	7	10.870	1.886
Mean	518	5	7.348	0.733
Std Dev	24	1	1.633	0.414

Table 5-3. Correlation Matrix of RT with Three Predictor Variables.

Variable	1	2	3	4
1. RT	1.00			
2. WL	0.11	1.00		
3. lnLF	-0.22	-0.33**	1.00	
4. CD	0.29*	0.19	-0.64**	1.00

* $p<0.05$ (1-tailed) ** $p<0.01$ (1-tailed)

CD and RT were significantly linearly related: Pearson $r=0.29$, $p<0.05$, one-tailed. As expected, lnLF was inversely related to RT: $r=-0.22$, $p=0.06$, one-tailed, but this was marginally significant. There was no linear relation between WL and RT, however ($r=0.11$, $p>0.10$). The limited range of the WL variable (see Table 5-2) is likely to blame. Although CD was correlated to the largest extent with response time, it was also intercorrelated with frequency ($r=-0.64$, $p<0.01$), indicating that frequency may be a confounding variable.

We next attempted to determine the unique relationship between CD and RT and between lnLF and RT using partial correlation techniques. Holding WL and lnLF constant, CD was still (marginally significantly) correlated with RT: $r=0.20$, $p=0.08$, one-tailed, but when variance shared with WL and CD was partialled out, no linear relation between lnLF and RT remained: $r=-0.03$. For this set of materials at least, CD appears to be the better predictor of word recognition performance.[8] A strong interpretation of the partial correlation results is that the zero-order correlation between RT and lnLF is largely spurious, due to the confound between lnLF and CD. We discuss the CD-frequency confound further in section 5.2.2 below.

The equation derived from a simultaneous multiple regression conducted on response latency using the three predictor variables is RT = 1.129 WL + -0.615 lnLF + 14.9 CD + 505.39. When WL and lnLF are included in the regression equation, RT

[8] Although it is not entirely valid to include the items used to optimise the parameters of the CD measure in the regression analysis, because of the problem of overfitting the data, we nevertheless carried out identical analyses over all the content words in the dataset ($n=123$). The zero order correlations were comparable: RT and CD: $r=0.38$, $p<0.0001$, one-tailed; RT and lnLF: $r=-0.35$, $p<0.01$; CD and lnLF: $r=-0.70$, $p<0.001$. There was no linear relation between RT and WL: $r=0.01$. The second-order partial correlations were also comparable: after controlling for WL and lnLF, CD and RT were still significantly correlated: $r=0.29$, $p<0.001$, but partialling out the effects of WL and CD left no linear relation between lnLF and RT: $r=0.00$.

is predicted to increase 14.9 ms for every bit of information conveyed by the target word about its contexts of use.

5.2.1.3 Discussion

The results of Experiment 9 have provided initial support for CD as a psychologically valid measure of lexical processing effort. A significant amount of response time variance was accounted for by the corpus-derived measure of the quantity of information conveyed by a word about its contexts of use. Words that appear in relatively constrained (or distinctive) contexts have high CD scores and tend to produce longer lexical decision latencies. In contrast, words whose contexts of use are unconstrained have low CD scores, and the time to classify them as real English words is shorter.

It should be emphasised that the linear relation observed between CD and lexical decision latency likely only holds when the target words are presented in isolation. With adequate contextual support, recovering the meaning of a high CD word should be no more difficult than for a low CD word, and consequently CD would not be expected to account for any significant amount of response time variance for words presented in context.

One avenue open to investigation is the choice of corpus used in the computation of CD. Does the distributional information contained in spoken language provide a more accurate model of language experience than the information present in written language? In order to address this question, we assembled a matched 10M word text corpus by randomly selecting texts from the 90M word written language portion of the BNC, and calculated CD using a ±5 word window and the 500 most frequent content words in this corpus as context words. Next, we conducted a linear regression analysis on the VLD latencies recorded for the entire set of 123 content words. Although $CD_{written}$ and CD_{spoken} were highly correlated (r=0.80, p<0.0001), CD determined from the matched text corpus turned out to be somewhat less predictive of response times than CD computed from the BNC-spoken: r=0.30, p<0.001, one-tailed for $CD_{written}$ compared with r=0.38, p<0.0001 for CD_{spoken}. Although this difference may be attributable to sampling error, a more attractive interpretation is that spoken language is the more psychologically relevant source of distributional information, reflecting the imbalance in one's exposure to the two language types in the environment.

5.2.2 CD and word frequency

Corpus frequency is an established, robust predictor of word recognition performance. High frequency words are recognised more quickly and with greater accuracy than low frequency words (see, *eg*. Monsell, 1991). Although the *word frequency effect* (WFE) has stood up well over the past few decades, a number of authors have contested its primacy, proposing that the effect is due to a spurious correlation of frequency with another, more legitimate variable, which has led to claims that variables such as Familiarity (Gernsbacher, 1984) and Age of Acquisition (*eg*. Morrison & Ellis, 1995) subsume frequency as predictors of performance in tasks such as lexical decision and naming. Similarly, the results of Experiment 9 suggest that the variance accounted for by frequency in tasks involving the recovery of word meaning may be better attributed to CD. This hypothesis has serious implications for the primacy of the WFE.

The subjective definition of CD as the informativeness of a word about its contexts of use anticipates its inverse relation with frequency. Frequent words tend to appear in a wider variety of contexts than infrequent words. Rare words, by definition, are not encountered very often at all; hence their contextual behaviour is naturally more constrained. We would therefore expect CD to increase as frequency decreases. This relationship was confirmed in Experiment 9, where a negative correlation (r=-0.64) was obtained between log-transformed lexeme frequency and CD. For comparison, the correlation coefficient for the set of 8,097 lexemes occurring 25 times or more in the BNC-spoken is -0.82.

Both CD and frequency are measurable properties of the linguistic environment, and both potentially influence lexical processing. Can we say much about the legitimacy of one variable over the other? The results of the partial correlation analyses reported above support CD as the stronger predictor, since no significant relationship between RT and frequency remained once CD was controlled. However, three objections to this conclusion can be raised. First, partial correlation analysis involving two strongly related variables favours the variable with the largest zero-order correlation with the variable of interest. Second, this conclusion is based on a single sample. The roles of CD and lnLF may well reverse when the same analysis is applied to a second sample of words. Replication of Experiment 9's pattern of results is clearly required in order to pursue an empirical distinction between the two lexical variables.

Table 5-4. Correlation Matrix of RT with Four Frequency Variables, for the Set of Words Used for Parameter Optimisation (n=70).

Variable	1	2	3	4	5
1. RT	1.00				
2. BNC-spoken	-0.43**	1.00			
3. BNC	-0.38**	0.91**	1.00		
4. CELEX-spoken	-0.38**	0.93**	0.92**	1.00	
5. CELEX	-0.41**	0.91**	0.97**	0.89**	1.00

* p<0.05 (1-tailed) ** p<0.01 (1-tailed)

Table 5-5. Correlation Matrix of RT with Four Frequency Variables, for Experiment 9's Materials (n=53).

Variable	1	2	3	4	5
1. RT	1.00				
2. BNC-spoken	-0.22	1.00			
3. BNC	-0.21	0.90**	1.00		
4. CELEX-spoken	-0.21	0.88**	0.90**	1.00	
5. CELEX	-0.25*	0.84**	0.93**	0.93**	1.00

* p<0.05 (1-tailed) ** p<0.01 (1-tailed)

The third potential objection against the strong view of CD subsuming frequency is that lexeme frequency derived from the spoken language component of the BNC is simply not a good model of frequency. The correlation we obtained between lexical decision latency and frequency is somewhat weaker than what other researchers have reported; however, the BNC-spoken is a much larger sample of language than the corpora on which other published frequency lists are based. In order to address this concern, we computed the correlation coefficients between RT and three other sources of (log-transformed) lexeme frequency: the entire 100 million word BNC, the CELEX database (17.9M words), and the spoken language counts from the CELEX database (based on 1.3M words).

Table 5-4 shows the correlations between each frequency estimator and the lexical decision latencies for the 70 words used for parameter optimisation. All frequency variables were highly intercorrelated; BNC-spoken frequency had the largest correlation with RT. The other three variables were comparable predictors, with correlation coefficients of around -0.40. For the 53 words used in Experiment 9, the CELEX count was a slightly better predictor of RT than the other three variables (Table 5-5), and was the only predictor to reach statistical significance:

r=-0.25, p<0.05, one-tailed. Because the four frequency variables are roughly equivalently predictive of RT across the two sets of data, it appears that objections regarding the adequacy of BNC-spoken lexeme counts as a measure of frequency are unfounded.

If we accept the conclusion that the response time variance explained by both CD and frequency is better attributed to CD, is there any role left for frequency in the current model of lexical processing? The formal definition of CD still requires frequency information: recall that the prior probability distribution is estimated directly from lexeme counts. We might suggest that frequency is useful for precisely this purpose. Frequency of occurrence provides an uninformative model of the contextual distribution of any word; given no other information (treating the corpus as an randomly ordered 'bag' of words), the estimated probability of a particular context word occurring in a window around any target word is simply the relative frequency of that context word.

5.2.3 Experiment 10

Although the results of Experiment 9 provide support for CD as a psychologically valid measure of the difficulty of processing a word in isolation, corpus frequency was a potential confounding variable. It is not yet clear if the roles of CD and frequency in accounting for human processing behaviour can be distinguished empirically. Experiment 10 is an attempt to establish a unique role for CD over word frequency, by explicitly manipulating CD in a visual lexical decision task with new materials and participants. Responses are predicted to be faster for words with low CD values than for high CD words. The existence of a basic contextual distinctiveness effect on VLD latencies should be apparent using two sets of words differing in their mean CD value. Furthermore, by closely matching pairs of stimuli on influential variables such as corpus frequency and word length, variability due to these factors can be substantially reduced, while permitting a more powerful repeated measures analysis of variance to be conducted by-items. If a CD effect is observed using a material set closely controlled for extraneous variables such as frequency, then this would be compelling evidence that CD has an independent influence on word recognition performance.

5.2.3.1 Method

Subjects. Twenty-four University of Edinburgh students were each paid £2 to participate. All were native speakers of English and had normal or corrected-to-normal vision.

Materials and design. The word stimuli consisted of 40 pairs of words individually matched for orthographic length and BNC-spoken lexeme frequency. Frequency ranged from 2.5 to 909 occurrences per million, and word length ranged from three to seven letters.

Word pairs were chosen such that the first member of each pair was substantially higher in CD than the second member. Selection of stimuli was also constrained by two other lexical properties, Age of Acquisition (AoA) and Familiarity (Fam), both of which have been argued to influence performance on word recognition tasks. AoA is a subjective estimate of the age at which one learns a word, and Fam is a subjective measure of previous experience: a rating of how familiar subjects are with a word (Gernsbacher, 1984). Normative AoA and Fam values for the critical stimuli were drawn from the MRC Psycholinguistic Database (Coltheart, 1981). As a result of the matching procedure, the set of High-CD words did not differ from the Low-CD group in terms of log-transformed lexeme frequency (paired $t[39]=0.34$, $p=0.74$), AoA (paired $t[39]=1.25$, $p=0.22$) or Fam (paired $t[39]=-0.80$, $p=0.43$). Mean CD for the High-CD words was 2.199, and the mean CD value for the Low-CD items was 1.391; this difference was highly reliable: paired $t(39)=22.49$, $p<0.001$.

The 80 nonword stimuli were *pseudohomophones* matching the word stimuli in letter length. Pseudohomophones such as *furst* and *krapht* are more word-like than typical nonwords used in lexical decision experiments; Joordens and Becker (1997) argue that their inclusion as foils in the lexical decision task forces responses to rely more heavily on semantic information, compared with using pronounceable nonwords. Forty-two of the pseudohomophones were taken from Joordens and Becker (1997), and the remainder were created anew. The list of 160 stimuli was randomised for each subject. Ten additional items (five words and five nonwords) were included to serve as a practice session. All materials are listed in Appendix H.

Procedure. Participants were tested individually. The PsyScope program (Cohen *et al.*, 1993) running on a Macintosh Performa 475 computer with 16" monitor

displayed the stimuli and timed responses. Each trial was initiated with a centered fixation point ('+') displayed for 500 ms, followed immediately by the letter string (in lowercase 18pt Palatino font).

Response times were measured from the onset of stimulus presentation until the subject pressed a key on the button box. The inter-trial interval was 250 ms. Failure to respond within 2000 ms also initiated the next trial.

Subjects were directed to respond as quickly and as accurately as possible, using their dominant hand to indicate if the letter string was a real English word, and their non-dominant hand to respond to a nonword. Following both verbal and written instructions and the 10 item practice session, the 160 stimuli were presented in two equal-sized blocks with a rest break before each block.

5.2.3.2 Results and Discussion

The 2000 ms response time limit was exceeded in only one 'word' trial; this was scored as an error. The overall error rate was low: 6.35% of real words were classified as nonwords.

Repeated measures analyses of variance on the error rate data did not reveal an effect of CD: $F_1(1,23)<1$, $F_2(1,39)<1$. There was also no CD effect on response times: $F_1(1,23)<1$, $F_2(1,39)<1$.

The multiple regression equation fit to Experiment 9's data predicted a unique influence of CD on VLD latency of 14.9 ms per bit. The difference between the mean High-CD and Low-CD values in the current experiment is 0.808 bits; this corresponds to an anticipated CD effect size of 12 ms. Why was this effect so elusive?

Upon closer examination of the response data, it was apparent that the spread of subjects' mean RTs was quite wide, suggesting that subjects could be legitimately separated according to their average response speed. It is possible that the faster subjects were utilising a qualitatively different strategy than slower subjects for the classification task, resulting in differential sensitivity to CD differences in the stimuli. If some subjects based their lexical decisions primarily on 'surface cues' (such as the visual familiarity of the orthographic form), which requires minimal semantic processing, their response latencies would tend to be shorter than those of subjects whose lexical decisions involved 'deeper', meaning-based processing. Differences in CD – a variable hypothesised to reflect the effort of recovering word meaning – may influence the VLD responses of slower subjects only.

Table 5-6. Mean VLD Response Times (in Milliseconds) and Error Percentages.

	Mean RT		Error Rate	
Group	High-CD	Low-CD	High-CD	Low-CD
By-subjects				
Fast	481	492	8.5	5.8
Slow	611	598	5.2	5.6
By-items				
Fast	487	496	8.5	5.8
Slow	613	601	5.2	5.6

Note: VLD=visual lexical decision; RT=response time.

This hypothesis was tested by carrying out a median split of the participants according to their mean RTs. There was a significant interaction between the Speed and CD factors: $F_1(1,22)=8.36$, $p<0.01$; $F_2(1,39)=1.89$, $p=0.177$. The Slow group's response latencies for High-CD items were, on average, 13 ms slower than for Low-CD items, consistent with the experimental hypothesis. In contrast, the Fast group responded 11 ms *faster* to the High-CD words (see Table 5-6). Although error rate differed between conditions for the Fast subjects (8.5% for High-CD words, compared with 5.8% for Low-CD words), indicating a potential speed-accuracy trade-off, the interaction between CD and Speed was not reliable: $F_1(1,22)=1.63$, $p=0.215$; $F_2(1,39)=3.70$, $p=0.062$.

In order to determine whether the lexical decision responses made by the two groups of participants were sensitive to different lexical properties, we computed the zero-order correlations between mean RT and five other variables besides CD: WL, lnLF, AoA, Fam, and *orthographic neighbourhood density* (see Table 5-7). Neighbourhood density (N) is a measure of orthographic similarity; it is calculated as the number of real words that can be created by changing one letter of the target word at a time.[9] N was included because it has also been claimed to predict word recognition performance (*eg.* Coltheart, Davelaar, Jonasson & Besner, 1977).

For the Fast group of participants, all predictor variables except for AoA and WL were significantly correlated with RT.[10] For the Slow group, all variables except

[9] We calculated N as the number of words in the CELEX database that are exactly one letter away from the target word.

[10] WL was *negatively* correlated with RT for the Fast group, which is opposite to the typical finding that response times tend to increase with word length. Because one-tailed tests were used, this correlation was not significant.

Table 5-7. Correlations Between Mean Response Times for the Fast (RT-Fast) and Slow (RT-Slow) Groups of Participants and Six Predictor Variables (n=80).

Variable	1	2	3	4	5	6	7	8
1. RT-Fast	1.00							
2. RT-Slow	0.45**	1.00						
3. WL	-0.33	-0.07	1.00					
4. N	0.49**	0.25*	-0.61**	1.00				
5. lnLF	-0.38**	-0.44**	0.13	-0.07	1.00			
6. AoA	0.13	-0.27**	0.29**	-0.21*	-0.08	1.00		
7. Fam	-0.55**	-0.55**	0.09	-0.12	0.60**	-0.51**	1.00	
8. CD	0.27**	0.45**	0.00	0.00	-0.73**	0.17	-0.53**	1.00

* p<0.05 (1-tailed) ** p<0.01 (1-tailed)

WL were reliably linearly related to RT. What is most interesting about these results is that the 'orthographic' variables, WL and N, were more strongly correlated with RT for the Fast group, and CD was more strongly correlated with RT for the Slow group. The pattern of results is consistent with the hypothesis that the faster subjects were processing the stimuli at a shallower, less 'semantic' level than the slower subjects.[11]

Since Fam had the strongest correlation with RT for both groups of participants (r=-0.55, p<0.01), it is possible that the ANOVA results were also influenced by the fact that stimulus pairs were not individually matched on this variable, even though High-CD and Low-CD group means were equivalent (see Appendix H). This was confirmed by correlation analyses conducted across items; there was a significant inverse relation between the *difference* in mean RT between the members of a matched pair and the *difference* in their Familiarity scores, for both the Fast and Slow groups of participants (r=-0.33, p<0.05; r=-0.54 p<0.001, respectively). If a High-CD word was more familiar than its matched Low-CD counterpart, it tended to elicit quicker lexical decision responses, and vice-versa. It appears that matching stimulus pairs individually on lnLF was not sufficient; the rated familiarity of a word had a substantial influence on how quickly it could be classified, independent of its corpus frequency.

[11] The tendency for the Fast group to respond *more* quickly in the High-CD condition is difficult to explain. A speed-accuracy tradeoff cannot completely be ruled out; the by-item analyses of variance hint that the Fast group was more error prone than the Slow group, F_2(1,39)=3.45, p=0.071, and that High-CD items elicited more errors than Low-CD items, but for the Fast subjects only (a CD × Speed interaction): F_2(1,39)=3.70, p=0.062. A difference between the two sets of items in terms of an uncontrolled 'orthographic' variable – to which the faster subjects are hypothesised to be more sensitive – is also a possibility.

Table 5-8. Summary of Individual Regression Analyses.

	Variable					
	WL	N	lnLF	Fam	AoA	CD
Fast group						
Mean *B*	-2.12	3.19	-4.63	-0.34	0.04	-5.19
SE	3.15	1.17	1.05	0.07	0.05	3.57
t(11)	-0.67	2.74	-4.41	-4.83	0.86	-1.45
2-tailed *p*	0.514	0.019	0.001	0.001	0.409	0.174
Slow group						
Mean *B*	6.46	3.92	-4.52	-0.28	0.07	12.28
SE	6.78	1.31	5.32	0.01	0.07	8.23
t(11)	0.95	2.98	-0.85	-2.02	0.90	1.49
2-tailed *p*	0.361	0.012	0.414	0.069	0.346	0.164

Although the correlation analyses have shown the CD measure to be a significant predictor of response latency for both groups of participants, it was confounded with lnLF and Fam (correlation coefficients of -0.73 and -0.53, respectively). We next attempted to clarify the unique influence of CD on RT, by conducting separate multiple regression analyses for each group of subjects. Regression analysis allows the influence of extraneous variables to be partialled out, in order to establish the unique role of the variable of interest.

Lorch and Myers (1990) have shown that it is inappropriate to average over subjects when carrying out multiple regression analyses on repeated measures data, because of the increased chance of making a Type I error when testing the significance of the partial regression coefficients. Since only item variability would be considered, the results of such a test could not be generalised beyond the sample of subjects. One remedy suggested by Lorch and Myers is to derive regression equations individually for each subject, and then test each regression coefficient to see if it is reliably different from zero using a one-sample *t* test. This was the approach taken here.

Before computing the two sets of 12 regression equations, lexical decision errors were replaced with the subject's mean RT. Table 5-8 summarises the results of the regression analyses. For the Fast group of subjects, there were reliable unique influences of N, lnLF and Fam after partialling out the variance shared with other variables in the equation. For the Slow group, there was a significant independent influence of N only. After controlling for the effects of the other five variables, no reliable unique relation between CD and VLD response latency remained, for either group.

In summary, the results of Experiment 10 have provided further support for the claim that CD models the effort of processing a word in the absence of context. Once the response time data were split according to average response latency, the CD effect was apparent for the slower group of subjects. The multiple regression analysis conducted on the Slow group's response data indicated that independently of the other variables, CD accounted for an increase in RT of 12.3 ms per bit, which was very close to the anticipated value. However, although this figure is an *accurate* estimate of the population regression coefficient (Lorch & Myers, 1990), it failed to reach significance when frequency was also included in the regression equation. This is an unavoidable consequence of collinearity: when two predictor variables are highly correlated, the standard errors of their partial regression coefficients are necessarily high, leading to low *t* values.

Two practical conclusions can be drawn from this experiment. First, the study has underlined the need for closely matching word stimuli on more lexical dimensions than corpus frequency; differences in familiarity proved to be correlated with RT differences, even though frequency was closely controlled on a pairwise basis. However, this approach to experimental design is severely limited in practice. As the number of variables on which stimuli are to be matched increases, it becomes extremely difficult, if not impossible, to find suitable items.

Second, it seems that the lexical decision task is less than ideal for testing claims about the influence of 'semantic' variables such as CD on the effort of recovering word meaning. The current subjects appeared to invoke qualitatively different strategies when confronted with the task of classifying letter strings; arguably, they were attending to different dimensions along which words vary. In order to measure the effort involved in recovering word meaning, a task is required that is both maximally sensitive to the semantic differences between words *and* discourages subjects from developing strategies that allow them to ignore these differences. It is not immediately apparent what kind of task would be appropriate; the semantic categorisation task – in which subjects respond by classifying words according to the presence or absence of a semantic property (such as animacy) – may meet these requirements.

5.3 Comparing CD with other semantic variables

In the word recognition literature a number of 'semantic' variables have been proposed to explain quantitative differences in lexical processing behaviour between otherwise matched words. Variables such as Concreteness are assumed to reflect between-word differences in semantic representation, or differences in how these representations are accessed. Since we have put forward CD as an information-theoretic measure of the processing effort involved in recovering word meaning, some insight may be gained into its nature by seeing how CD compares to these other 'semantic' variables. CD differs in at least one crucial respect; it is *objective* – being derived from a corpus of natural language – whereas variables such as Ambiguity/Polysemy, Concreteness and Imageability are *subjective* – they are measured using ratings elicited from human judges.[12]

Concreteness has been the most intensively studied of the lexical properties thought to influence semantic processing (*eg*. James, 1975; Paivio, 1971; Schwanenflugel & Shoben, 1983). Concrete words have referents which can be perceived by the senses (*eg*. *spoon*), and are responded to more rapidly than abstract words (*eg*. *favour*) in isolated word recognition tasks such as lexical decision and naming (*eg*. James, 1975; Schwanenflugel, Harnishfeger & Stowe, 1988). The concrete word advantage – really an abstract word disadvantage – is assumed to reflect the processing difficulty in retrieving the meaning of an abstract word out of context (abstract words rely more heavily on context for interpretation). This difficulty is considered to correspond to representational differences between the two types of words: concrete words are assumed to be represented, on average, more independently from their semantic contexts than abstract words, and Schwanenflugel and Shoben (1983) suggest that "there is generally more information associated with abstract concepts than concrete concepts" (p. 93). The claim that Concreteness crucially affects *semantic* processing is supported by studies of bilinguals (*eg*. de Groot, 1992), which find an advantage for concrete over abstract words in translation tasks. Concrete words are thought to have more commonality in meaning representation across languages than abstract words.

Context Availability (CA) is a subjective measure of the ease with which one can think of a particular circumstance in which a word might appear, and has been

[12] Even though Ambiguity-type scores are typically calculated from dictionary entries (*eg*. Jastrzembski, 1981), and the creation of modern dictionaries has been assisted by computational corpus analysis, they are ultimately based on the intuitions of the lexicographers.

argued to be the 'real' explanatory variable underlying concreteness effects (Schwanenflugel & Shoben, 1983). Concreteness is highly correlated with CA, since it is generally easier to think of a context for concrete words than for abstract words. The concrete word advantage in lexical processing may simply reflect the fact that contextual information for abstract words is more difficult to retrieve. In order to test this hypothesis, Schwanenflugel and Shoben (1983, Experiment 3) used partial correlation techniques to assess the contribution of Concreteness in accounting for lexical decision response variance, when CA was held constant. Concreteness was no longer a significant predictor with CA partialled out; conversely when Concreteness was held constant, CA was still significantly correlated with response latency. Because Schwanenflugel and Shoben have demonstrated the existence of Concreteness effects when target words are presented in isolation, but not when following a sentence context, Context Availability seems to be the more theoretically attractive variable, since there is a clear role for prior context to affect word recognition. If retrieval of a word's contextual information is not required, as in the case when this information is available from the immediate linguistic context, then processing of abstract words should be no more difficult than the processing of concrete words, all else being equal.

The third 'semantic' variable to be examined is Ambiguity (also known as Polysemy and Number of Meanings). Lexical decisions made to ambiguous words presented in isolation are faster than to matched unambiguous items (*eg*. Jastrzembski, 1981). The Ambiguity advantage lacks a compelling explanation; early work considered the effect to be a consequence of lexical access: the more 'meaning entries' in the mental lexicon that correspond to a word form, the more rapidly that word can be recognised.

In Experiments 11 and 12, we use correlation analysis to examine the relationship between CD and the three 'semantic' variables introduced above (Concreteness, Context Availability, and Ambiguity), in order to investigate three hypotheses. The first prediction is that that concrete words can be differentiated from matched abstract words by their CD values. If Schwanenflugel and Shoben's (1983) claim that more contextual information is associated with abstract words than with concrete words is correct – and assuming that CD captures their notion of 'contextual information' – we should find an inverse relationship between Concreteness and CD. The amount of information conveyed about a word's contexts of use should, in general, be greater for abstract than for concrete words.

Since Context Availability and CD intuitively measure closely related lexical properties – the one essential difference being that CA ratings are subjective, whereas CD values are objective – the ease with which one can think of a context for a word should be inversely related to the distinctiveness of the contexts in which it appears. Words which have high CA ratings (such as *office*) should tend to have relatively unconstrained contexts of use, and thus low CD scores.

The third hypothesis concerns the relationship between CD and Ambiguity. Words described as lexically ambiguous, *ie.* words that are used to communicate a number of meanings, intuitively would occur in a wide range of contexts. Thus, words with high Ambiguity scores should tend to be less contextually distinctive than minimally ambiguous words, and consequently should convey less information about their contexts of use. CD is predicted to be negatively correlated with Ambiguity. In Experiment 13, we also compare CD with Ambiguity in terms of behavioural prediction, through reanalysis of a recent word recognition study.

5.3.1 Experiment 11

5.3.1.1 Method

Concreteness and Context Availability ratings for 160 words were drawn from norms collected by Tokowicz (1997, Experiment 2). This set of words was originally compiled as materials for a series of experiments examining models of bilingual lexical representation. Words were rated for Concreteness and CA by independent groups of 13 and 11 subjects, respectively, and were divided into two groups of 80: abstract and concrete, individually matched for frequency and word length. Five words in the abstract group had an BNC-spoken frequency of less than 25 occurrences; we excluded these words and five frequency-matched concrete words from further analysis. After removal of these items, the two groups were still reasonably matched on the relevant variables. For the concrete group (mean Concreteness score of 5.96) the values were as follows, Context Availability: 6.35; log-transformed lexeme frequency (lnLF): 5.957. For the abstract words (mean Concreteness of 3.75), Context Availability: 5.79; lnLF: 5.850.

Ambiguity scores for the 150 stimuli were computed as the number of senses associated with each word's entry in the WordNet lexical database (Miller *et al.*, 1990). The number of senses corresponding to a WordNet entry is roughly comparable to the number of 'meanings' in a dictionary definition, since WordNet

Table 5-9. Rank Correlations Between CD and Three 'Semantic' Variables.

Variable	1	2	3	4
1. CD	1.00			
2. CA	-0.22**	1.00		
3. Concreteness	-0.11	0.75**	1.00	
4. Ambiguity	-0.38**	-0.10	-0.13	1.00

* $p<0.05$ (1-tailed) ** $p<0.01$ (1-tailed)

was created using lexicographic methods. Ambiguity values were logarithmically-transformed to reduce skew in the distribution.

5.3.1.2 Results and Discussion

Visual examination of the distributions for several of the variables indicated that non-parametric statistical tests would be appropriate. Table 5-9 displays the rank correlation matrix for the four variables. As expected, Concreteness and Context Availability were highly correlated (Spearman $\rho=0.75$, $p<0.01$, one-tailed. CD was negatively correlated with Concreteness ($\rho=-0.11$, $p=0.09$, one-tailed), but this was only marginally significant. However, a Mann-Whitney *U* test revealed the mean CD value of the concrete group of words ($M=1.04$, $SD=0.55$) to be significantly smaller than the abstract group ($M=1.32$, $SD=0.78$): $U=2223$, $z=-2.216$, $p<0.05$. Although there was no reliable association between CD and Concreteness, the hypothesis that concrete words can be distinguished from matched abstract words in terms of contextual distinctiveness was confirmed. Assuming all other factors to be equal, the set of concrete words conveyed less information about their contexts of use than the set of abstract words.

CD was significantly negatively correlated with Context Availability ($\rho=-0.22$, $p<0.01$, one-tailed), indicating that the more constrained a word's linguistic contexts of use, the more difficult it was for subjects to think of a context for that word.

CD was also negatively correlated with Ambiguity ($\rho=-0.38$, $p<0.01$, one-tailed), supporting the hypothesis that words that tend to appear in relatively constrained contexts (and thus have high CD values) are less ambiguous than words that occur in less constrained contexts.

5.3.2 Experiment 12

The purpose of Experiment 12 was to examine the relationship between CD and the same three 'semantic' variables (Concreteness, CA and Ambiguity), across a different set of words. These were drawn from the 64 nouns used by Schwanenflugel and Shoben (1983) in their study of concreteness effects. The authors included a fourth subjective variable in their regression analyses: Number of Contexts. In order to elicit judgements for this variable, subjects were instructed that "some words appear in a greater variety of contexts or as part of a description of a greater number of situations than other words," and so were to judge "the relative diversity of contexts" for each word, using a 9-point scale, where a 1 indicated that the word appears in a minimal number of different contexts, and a 9 meant it appears in a large number of contexts. Number of Contexts would appear to share one aspect of lexical variation captured by CD: the greater a word's rated "diversity of contexts", the less constrained its contextual behaviour. CD is predicted to be inversely related to Number of Contexts.

5.3.2.1 Method

The method for Experiment 12 was identical to Experiment 11, except for the difference in materials. The 64 noun stimuli were pre-divided by Schwanenflugel and Shoben (1983) into two equal-sized groups of abstract and concrete words, matched for word length and frequency, and were rated for Concreteness and Context Availability by independent groups of 20 and 22 subjects, respectively. The latter group also provided Number of Contexts ratings. Of the 64 items, one concrete and four abstract words failed to make the frequency threshold of 25 occurrences in the BNC-spoken, and consequently these words and their five matched counterparts were not included in the analyses.[13]

5.3.2.2 Results and Discussion

The rank correlations between CD and the other four 'semantic' variables are displayed in Table 5-10. Although the size of Experiment 12's material set is substantially smaller than Experiment 11's, a useful comparison of the results of the correlation analyses is still possible.

[13] 15 items were originally listed in their plural form; these words were first converted to their singular (canonical) form in order to calculate CD.

Table 5-10. Rank Correlations Between CD and Four 'Semantic' Variables.

Variable	1	2	3	4	5
1. CD	1.00				
2. CA	0.04	1.00			
3. Concreteness	-0.20	0.76**	1.00		
4. Ambiguity	-0.39**	-0.24*	-0.11	1.00	
5. NumContexts	-0.10	-0.72**	-0.48**	0.32*	1.00

* $p<0.05$ (1-tailed) ** $p<0.01$ (1-tailed)

Paralleling Experiment 11's findings, Concreteness and Context Availability were highly correlated (Spearman $\rho=0.76$, $p<0.01$, one-tailed). CD was negatively correlated with Concreteness ($\rho=-0.20$, $p=0.077$, one-tailed), which provides some support for the hypothesis that concrete words tend to occur in less constrained contexts than abstract words. The higher a word's rated Concreteness value, the less information that word tends to convey about its contexts of use. However, contrary to Experiment 11, there was no significant difference in the mean CD value for the concrete word group ($M=0.89$, $SD=0.38$) compared to the abstract group ($M=1.06$, $SD=0.59$): Mann-Whitney $U=321$, $z=-0.75$, $p=0.45$.

In contrast to the results of Experiment 11, there was no relationship between CD and CA ($\rho=0.04$).

CD was negatively correlated with Ambiguity ($\rho=-0.39$, $p<0.01$, one-tailed), which was in accordance with the results of Experiment 11: the more information a word conveys about its contexts of use, the fewer lexicographically-defined senses (or meanings) the word has.

Number of Contexts was not significantly correlated with CD ($\rho=-0.10$), which indicates that CD cannot be straightforwardly interpreted as a measure of the variety of contexts a word occurs in (assuming, of course, that Number of Contexts is indeed a valid measure of this quantity). Number of Contexts did correlate with Concreteness ($\rho=-0.48$, $p<0.01$) and CA ($\rho=-0.72$, $p<0.01$). Although abstract words were judged to appear in more diverse contexts than concrete words, the subjective Number of Contexts variable was *inversely* related to CA. For some reason, if a word is judged to appear in a wide variety of contexts, it tends to be more difficulty to bring these contexts to mind (*ie.* they are less 'available') when CA ratings for the word are elicited.

5.3.3 Experiment 13

The inverse relationship between CD and Ambiguity noted in Experiments 11 and 12 deserves further exploration. From these studies, it is clear that ambiguous words tend to be less contextually distinctive than unambiguous words (at least when Ambiguity is operationalised as number of WordNet senses). CD and Ambiguity were also significantly correlated for the 53 words analysed in Experiment 9: r=-0.48, p<0.001, and for the 80 items in Experiment 10: r=-0.40, p<0.001. The existence of this reliable negative correlation raises the possibility of a confound between the two variables – Ambiguity effects found using word recognition tasks might be better explained as CD effects.

In a similar vein, it is also possible that some experimental studies have *failed* to reveal a facilitatory effect of Ambiguity on lexical processing behaviour because of the confound with CD. In other words, two groups of stimuli might differ reliably in Ambiguity, but they might not differ in CD. We tested the hypothesis that a CD confound could eliminate a potential Ambiguity effect by calculating the difference in CD between two carefully matched groups of words differing in their mean number of meanings. The materials from Borowsky and Masson (1996, Experiments 2 & 3) were suitable for this purpose.

5.3.3.1 Method

Borowsky and Masson compiled a list of 64 ambiguous words from previous published research, and selected 64 unambiguous words closely matching on such variables as word length, corpus frequency and orthographic neighbourhood density. An unambiguous item was defined as such by having a single entry in a paper dictionary.

Five of the 64 pairs contained an inflected word form; these pairs were discarded. Of the remaining 118 words, seven ambiguous and 13 unambiguous items were too low in corpus frequency (occurring less than 25 times in the BNC-spoken) to calculate CD reliably, requiring the removal of 16 more pairs from the materials. A total of 43 pairs of matched items remained for the reanalysis. The two groups were still reasonably close in terms of log-transformed lexeme frequency (ambiguous group: 5.879; unambiguous group: 5.797), even though Borowsky and Masson originally matched items for surface word form frequency. In order to confirm their classification of the stimuli in terms of ambiguity, Ambiguity scores were computed for each item from WordNet. The ambiguous group was significantly

higher in WordNet-derived Ambiguity: paired $t(42)=3.61$, $p<0.001$, one-tailed. The ambiguous group had an average of 12.5 WordNet senses, compared with 7.7 senses for the unambiguous group.

5.3.3.2 Results and Discussion

CD was determined for each item using the same procedure followed in Experiments 9-12. Consistent with the negative correlation obtained between CD and Ambiguity for the materials of Experiments 9-12, CD and Ambiguity were again significantly inversely related: Pearson $r=-0.38$, $p<0.001$. Mean CD for the ambiguous items was 1.105 ($SD=0.514$) and 1.094 ($SD=0.495$) for the unambiguous words. The difference in CD was not significant: paired $t(42)=0.12$.

The present results suggest an alternative interpretation of the experimental results reported by Borowsky and Masson (1996). The ambiguity advantage sought by Borowsky and Masson was observed in only one of their four experiments, and the effect was reliable in the by-subjects analysis of variance only. Standard naming tasks and lexical decision using orthographically illegal nonwords did not reveal an ambiguity effect. Although the two groups of words compiled from Borowsky and Masson's materials were distinguishable in degree of ambiguity (verified by the reliable difference in our WordNet-derived Ambiguity scores), they were not distinguishable in terms of CD.

This reanalysis offers some support for CD as a more relevant determinant of lexical processing effort than an Ambiguity-type variable. The inconsistently observed effect of Ambiguity in the word recognition literature may be due to the presence of CD as a confounding variable. Our explanation for why the expected semantic ambiguity effect was not observed by Borowsky and Masson is that their two groups of stimuli did not differ in the relevant variable, CD. They did differ in Ambiguity, a variable that we have found to correlate with CD, but which is simply not the critical lexical property affecting processing behaviour.

5.3.4 General discussion

In summary, correlation analyses pitting CD against several 'semantic' variables proposed in the word recognition literature have indicated a less than clear-cut pattern of results. Although CD was consistently correlated with Ambiguity, the presence or absence of reliable associations with Concreteness and Context Availability was dependent on the set of materials used. There were marginally

significant correlations between CD and Concreteness in Experiments 11 and 12, but only in Experiment 11 was there a significant inverse relationship between CD and CA. It appears that Concreteness and Context Availability do not draw upon the same source of information that CD does.

Experiments 11-13 (together with Experiments 9 and 10 above) were consistent, however, in demonstrating a relationship between CD and Ambiguity – CD was significantly negatively correlated with Ambiguity across five different sets of words. This finding raises the possibility that reported effects of Ambiguity-type variables on word recognition performance may be better attributed to CD. Experiment 13 provided indirect evidence for this claim; two groups of stimuli differing in Ambiguity but matched on a host of other variables did not produce a reliable Ambiguity effect. The fact that these two groups did not differ in CD suggests that CD is the more relevant variable for predicting lexical processing effort.

Taken together, Experiments 9-13 have provided evidence for the psychological reality of a new lexical property. Contextual distinctiveness has been shown to be a dimension of lexical variation that is relevant to human language processing behaviour, which future psycholinguistic research should recognise. Experimental designs that rely on matched lexical stimuli will need to take into account the potential confounding effect of CD.

5.4 CD and semantic impairment in Alzheimer's dementia

So far in this chapter, we have put forward CD as a variable that reflects the effort involved in recovering the meaning of an isolated word. This proposal has received some support from the word recognition studies (Experiments 9 and 10) reported earlier, which investigated on-line lexical processing behaviour. Does CD additionally offer a plausible explanation of *off-line* semantic phenomena? A further test of the psychological validity of CD is prediction of the types of words which cause the most difficulty for persons with certain kinds of language impairment. For example, anomia (impaired naming ability) is a common language deficit for people with dementia of the Alzheimer's type (DAT), and its severity can be assessed through picture naming tasks. Performance on the picture naming task is thought to depend heavily on semantic processing, and thus DAT patients' naming ability

offers a further testing ground for the current view of CD as a measure of processing effort.

Chenery, Murdoch and Ingram (1996) propose that anomia in DAT is due to a deficit in semantic processing. They administered the Boston Naming Test (BNT; Kaplan, Goodglass & Weintraub, 1983) to 23 people diagnosed with DAT and 23 matched control subjects. The BNT requires subjects to provide a name for 60 simple pictures; the target names are all concrete nouns. Response errors were classified into four major types: semantic, perceptual, phonological and no-relationship. Chenery *et al.* were interested in the relation between the severity of the dementia and the type and number of errors made, and thus did not analyse naming performance on a *by-item* basis. However, it is possible that certain target names were more vulnerable to impairment than others.

Chenery *et al.* found that as the severity of dementia increased from the mild to the moderate stage, both semantic and no-relationship errors increased markedly. From their analysis of the semantic error subtypes, they suggest that impairment is located in the retrieval mechanism, as opposed to reflecting problems with the structure of semantic memory, since responses often had a recognisable semantic relation to the target name. No-relationship errors arose when the subject was unable to retrieve *any* semantic information about the object name. In contrast, perceptual and phonological errors, which are assumed to have a non-semantic basis, generally occurred much less frequently than the two other types of error.

On the basis of a detailed analysis of the naming errors made by their patients, Chenery *et al.* propose that the semantic processing deficit in DAT is primarily due to a breakdown in the mechanism for accessing semantic representations (at least in DAT of mild to moderate severity). If this claim is correct, we can formulate two CD-based hypotheses. First, we predict that object names whose meanings are difficult to retrieve are named correctly by a smaller proportion of DAT patients than words whose meanings are more easily accessed. If, as we have argued in this chapter, the effort in recovering the meaning of a word is predicted by the amount of information that it conveys about its contexts of use, then high CD words should be more impaired than low CD words.

A second hypothesis concerns the potential for an interaction between DAT severity and CD. Are high CD object names more impaired than low CD names in the early stages of DAT? If anomia is initially observed for words whose meanings are the most difficult to recover, then we might expect the severity of impairment

and CD to interact; specifically, low CD names should elicit a larger proportion of correct responses than high CD names from patients with a mild level of dementia. With progression of the disease, the difference in performance between low and high CD words should decrease, because the integrity of semantic memory in general (not only the access mechanism) is assumed to be compromised in patients with a severe level of dementia.

A straightforward approach to testing the first hypothesis is to determine whether CD is linearly related to the proportion of patients naming an object correctly. Detailed response data for the 23 DAT patients was kindly provided by Helen Chenery. Unfortunately, about half of the BNT materials were of very low frequency, meaning that reliable CD values could not be obtained from the 10M word BNC-spoken. Consequently, we decided to use the entire 100M word BNC to calculate CD.[14] However, even with this increase in corpus size of an order of magnitude, three picture names still occurred extremely rarely,[15] and so were excluded from further analysis.

A simple linear regression showed an item's CD value to be inversely related to the proportion of the DAT patients (collapsing together the Mild, Moderate, and Moderate/Severe groups) who correctly named that item: Pearson r=-0.47, p<0.001. As the target name's CD value increased, naming performance tended to decrease; this relationship confirms the first hypothesis.

The second hypothesis was addressed using a 3×2 factorial design, with Item as the random factor and Severity (Mild, Moderate, Moderate/Severe) and CD (High, Low) as fixed factors. (The median CD value was used to divide the stimuli into High-CD and Low-CD groups of 28 words each.) The High-CD group had a mean CD value of 1.864 bits, and the Low-CD group had a mean CD score of 0.263.

An ANOVA revealed a main effect of CD. Low-CD items were named correctly by a larger proportion of subjects than were High-CD items: $F(1,54)=20.49$, $p<0.001$ (see Table 5-11). This is consistent with the results of the regression analysis.

There was also a main effect of Severity, $F(2,108)=23.16$, $p<0.001$, indicating that the proportion of patients making correct responses differed between Severity conditions. The best performance was achieved by the group diagnosed with a Mild level of impairment (M=0.633, SD=0.275), and the worst performance was by the

[14] The 500 most frequent content words in the 100M word corpus served as context words, and co-occurrence was defined using a window size of ±5 words.

[15] These were *pretzel*, *seahorse* and *sphynx*, which had lexeme frequencies of 10, 18 and 13, respectively.

Table 5-11. Mean Proportion of Subjects Responding Correctly (with Standard Deviations), by Severity Group.

Severity	Low-CD		High-CD		Difference
	M	SD	M	SD	(Low – High)
Mild	0.770	0.253	0.495	0.225	0.275
Moderate	0.648	0.276	0.291	0.253	0.357
Mod/Sev	0.564	0.353	0.341	0.222	0.223

Moderate/Severe group (M=0.452, SD=0.313). These results are in agreement with the by-subjects analysis of variance conducted by Chenery *et al.*

The interaction between CD and Severity was marginally significant: $F(2,108)=2.70$, $p=0.072$. The difference in the proportion of patients correctly naming High and Low CD items was larger for both the Mild and Moderate groups than the Moderate/Severe group. Roughly speaking, as diagnosed Severity increased, the CD effect became (marginally) significantly smaller. Thus, the second hypothesis is supported: words whose meanings are generally difficult to retrieve (High-CD) were notably more impaired than Low-CD words in the early stages of DAT. This difference was less pronounced for the Moderate/Severe level of impairment, which is consistent with Chenery *et al.*'s account of the advanced stages of DAT involving a general breakdown of semantic function.

In summary, CD has provided a quantitative account of processing difficulty on a by-item basis, which suggests an information-oriented interpretation of the semantic impairment typical of Alzheimer's dementia. Picture naming difficulty was predicted to be greatest for objects whose names convey the most information about their contexts of use. This prediction was confirmed by the significant inverse relation obtained between CD and naming performance.

Although a deficit in the retrieval of word meaning is a compelling view of anomia in DAT, it cannot be the whole story. Nebes (1989) emphasises the task-dependent nature of semantic impairment:

> … it is unlikely that one can ascribe all semantic problems in [Alzheimer's] to a retrieval deficit, because the performance of demented patients can vary greatly on semantic tasks that make apparently similar demands on retrieval (Nebes, 1989, p. 390).

In addition, Nebes states that it is extremely difficult to determine the source of semantic impairment; even if performance is shown to vary inversely with the

demands of the retrieval task, this could nevertheless reflect a combination of semantic memory loss *and* difficulty in accessing semantic information.

Finally, the frequency (or familiarity) of the object name is an obvious alternative explanation for the results of our by-item analysis. The less frequent the target name, the greater the difficulty in naming the object (see, *eg.* Kirshner, Webb & Kelly, 1984). There is some evidence, however, that favours CD over a frequency-based explanation of DAT patients' naming performance. Nebes, Boller and Holland (1986) report an effect of *contextual constraint* on the ability of DAT patients to perform a sentence completion (cloze) task. The time to generate a completion word for a sentence fragment showed the same pattern as a group of normal old subjects; performance was near-normal when the context was highly constraining.[16] There is no obvious explanation for this finding if naming ability is closely tied to word frequency, which is typically construed as a lexical level (non-semantic) property. The CD measure, however, can easily incorporate the constraining effect of context. If the prior distribution – which models the processor's expectations about word meaning – is adjusted to incorporate the semantic constraints imposed by the preceding sentential context, then less information will be conveyed when retrieving the meaning of an upcoming word whose meaning matches these expectations. Target words which form sensible (*ie.* semantically plausible) sentence endings are predicted to be easier to integrate into the context than targets which are not sensible completions. This idea is formalised in Chapter 6.

5.5 Summary

This chapter has focussed on the development and testing of a novel information-theoretic measure of environmental complexity, which is derived from a word's co-occurrence representation. We presented this measure, *contextual distinctiveness* (CD), as a way to formalise the expectation-building strategy proposed in Chapter 1, and described how CD captures distributional differences between words. Experiments 9 and 10 revealed CD as a predictor of the effort involved in processing isolated words, and illustrated the natural confound between CD and

[16] In a related experiment, Nebes and Brady (1991) observed an analogous effect of contextual constraint using a comprehension task. The time taken for DAT subjects to judge whether a target word could sensibly complete a sentence fragment also varied with the amount of contextual constraint, and the pattern of performance was again similar to that of the normal old.

corpus frequency. Although suggestive at present, the existence of this confound invites reinterpretation of the word frequency effect. We then compared CD with other 'semantic' properties of words such as concreteness and ambiguity (Experiments 11-13), and concluded that CD provides a robust, objective measure of lexical ambiguity. Finally, we investigated CD's ability to predict the degree of semantic impairment reported in a study of picture naming performance by Alzheimer's patients. The analysis demonstrated that CD is able to provide a quantitative account of the lexical processing difficulty observed in an impaired population, on a by-item basis.

6.
Lexical Processing in Context

Current theories of word recognition lack convincing accounts of how the influence of context on processing behaviour should be accommodated. This chapter proposes that by updating the prior knowledge component of the information-theoretic model of lexical processing effort developed in Chapter 5, the effect of previous context can be integrated quite naturally. We begin by presenting evidence for the *incremental* nature of comprehension, and suggest that a parsimonious model should implement incrementality. Next, we formally describe a Bayesian technique for incorporating context, and assess the psychological validity of the incremental model through computational reanalysis of a single-word lexical priming experiment. We then reanalyse two eye movement studies that manipulate contextual constraint, and test the model against a corpus of eye movement data. Finally, we conclude by arguing that the proposed model provides a unified account of semantic context effects, as well as supplying a parsimonious explanation of lexical processing effort, in the presence or the absence of a constraining semantic context.

6.1 The incremental nature of semantic interpretation

Spoken language comprehension is by necessity a *temporal* process. Utterances are produced over time, and it seems clear that understanding also proceeds temporally (*eg*. Marslen-Wilson & Tyler, 1980). *Incremental interpretation* refers to the rapid, step-by-step integration of the meaning of successively encountered words into the

semantic representation of the entire utterance.[1] The purpose of this section is to underline the importance of incorporating the notion of incrementality into a model of lexical processing, and to describe the construction of such a model using the statistical framework already developed in previous chapters.

The incremental view of language comprehension suggests that a semantic interpretation is available at any point during the processing of an utterance, which is attractive for rapid resolution of syntactic ambiguity (Steedman, 1989). There is a second appealing functional motivation for incrementality: accessibility of partial meaning of an utterance before the end of some syntactic unit such as a clause or sentence is reached means that the processor is more easily able to cope with 'noisy' input (due to auditory interference, disfluency, etc.). If the speaker's communicative intent can be recovered, at least in part, without reliance on a complete parse, then transmission of information from speaker to hearer will be more efficient.

It is clear that a parsimonious theory of lexical processing needs to adequately account for the incremental nature of comprehension. Besides the wealth of evidence showing that the semantic properties of the previous context have a substantial influence on word recognition processes (*cf.* sections 4.2 and 4.3), there are also experimental data demonstrating the incremental nature of this influence on on-line processing. In an event-related brain potential (ERP) study, Van Berkum, Hagoort and Brown (1998) found very rapid disruptive effects for words that were semantically anomalous in either the sentence or discourse context, suggesting that the meanings of individual words are integrated into a discourse-level representation as the input unfolds. Additional evidence is provided by an investigation of syntactic ambiguity using the self-paced reading task (Altmann & Steedman, 1988, Experiment 2); this study indicated that semantic interpretation does not wait until a sentence or clause boundary is reached. Finally, indirect support for incrementality is provided by sentence priming studies where syntactic coherence is manipulated: context effects either diminish or disappear completely when the words in the sentence context are randomly ordered (*eg.* Masson, 1986). In order to build up a semantic representation for a complete sentence (or utterance), it appears that the ability to construct coherent partial representations at intermediate

[1] The semantic representation of a complete utterance must obviously integrate information from a number of sources besides the meanings of its constituent words, such as knowledge about the world, as well as incorporating constraints from syntax and pragmatic plausibility. We use the term *semantic representation* in a much narrower sense, to signify the representation of meaning derived only from the words in the utterance.

points is also required. The meaning of a sentence is not simply the "sum of its parts".

In this chapter we show that it is a straightforward matter to simulate incremental interpretation using the information-theoretic measure of contextual distinctiveness (CD) developed in Chapter 5. CD can easily be modified to implement the 'conditioning' effect of the preceding linguistic context.

Recall that the prior distribution component of the CD measure represents the processor's expectations about word meaning, and CD is an estimate of the difference between the expected meaning of a word and its 'actual' meaning. If the prior distribution is iteratively revised to take into account the new evidence provided by each successive word[2] in an utterance, then estimates of processing effort become available on a word-by-word basis. Thus, the prior distribution, in a broad sense, models the incremental construction of utterance meaning as a sequence of iteratively-refined expectations about the contextual behaviour, and hence the meaning, of upcoming words. Corresponding to Steedman's (1989) hypothesis that partial semantic *interpretations* are accessible before the end of an utterance, under the theory that comprehension involves a process of expectation-building, *expectations* about the meaning of upcoming words would be available at every point in the input. In the following sections we develop this idea further, by first formalising the updating of the CD measure's prior distribution, motivating the settings of two free parameters, and finally validating the model against lexical priming data.

6.1.1 The ICD model

In this section, we describe the approach taken for modifying Chapter 5's measure of contextual distinctiveness in order to take into consideration the previous linguistic context. We define *Incremental CD* (ICD) as the quantity of information provided by word *w* about its contexts of use, given that some linguistic context cue *c* has just been encountered. We call our computational-level model of lexical processing effort the *ICD model*. It is referred to as *incremental* because the amount of information conveyed by *w* now depends on the specific context it occurs in, and is

[2] More precisely, each successive *content* word: we assume that function words do not contribute to the formation of semantic expectations (although they are predictive of grammatical properties such as syntactic category), and therefore will not influence the form of the prior distribution.

also a function of the order of the words in this context. For convenience, we use *ICD* to distinguish the quantity of information conveyed by a word in context from *CD*, the amount of information conveyed by a word in isolation, even though they are exactly the same information-theoretic measure (relative entropy). The crucial difference is that CD employs an uninformative, relative frequency-based prior distribution, whereas the ICD model uses a prior derived from the contextual distributions of the words in the previous context.

The previous context, even a single word, is often informative about the identity, and consequently, the meaning of *w*. For example, the adverb *hermetically* is a very strong cue that the next word will be *sealed*,[3] and thus it is also a very strong cue that the *meaning* of the next word has something to do with a state of closure. The prior knowledge that the processor has about the meaning of an upcoming word should influence the amount of information conveyed upon observation of that word. Intuitively, less information about the meaning of *sealed* should be conveyed when it is encountered immediately after *hermetically*, compared with the case where *sealed* is preceded by a neutral, or less informative cue such as *carefully*. This prior knowledge about the meaning of *w* we term the *semantic expectation* – it is a type of prediction about the meaning of *w*. We shall now formalise a computational procedure for updating the prior information available to the processor after encountering a single context cue *c*, and then show how ICD values are calculated.

We use Bayes' theorem to revise the CD measure's prior probability distribution on the basis of the new evidence provided by *c*. The updating of the processor's expectations about meaning can be expressed as a Bayesian update rule, where the revised semantic expectation depends on the previous expectation and the likelihood of the context cue. More formally, the posterior (or updated) probability of the contextual distribution θ given *c* is proportional to the likelihood of observing *c* given the contextual distribution θ, multiplied by the prior probability of θ:

$$P(\theta|c) = \frac{P(c|\theta)P(\theta)}{P(c)} = \frac{P(c|\theta)P(\theta)}{\int_{-\infty}^{\infty} P(c|\theta)P(\theta)d\theta}$$

Although this is exactly what we need, this posterior density is often extremely difficult to evaluate. Fortunately, a computationally tractable solution is available

[3] This statement is supported by corpus evidence: *hermetically* occurs 23 times in the 100 million word BNC; in 20 of these instances it is immediately followed by *sealed*.

which capitalises on the fact that by choosing a prior distribution that is a *conjugate family* for the likelihood function, the posterior distribution is automatically expressed using the same functional form. In order to take advantage of this computational convenience, we first outline the central assumption about co-occurrence vectors required to implement this method.

A co-occurrence vector can be considered to represent the outcome of a *multinomial experiment*, if each context in which word *w* occurs is treated as an independent multinomial trial. So, in each of *n* 'trials' there are *k* possible outcomes, where *k* is the number of context words, and the probability of a particular outcome *i* is θ_i.[4] The multinomial distribution is fully specified by the set of parameters $\theta = (\theta_1 \ldots \theta_k)$, which sum to 1, and is written as:

$$P(c_1,...,c_k|\theta) = \frac{n!}{c_1!\ldots c_k!}\prod_{i=1}^{k}\theta_i^{c_i}$$

This equation expresses the *likelihood* of a particular co-occurrence vector $(c_1 \ldots c_k)$ arising from random sampling from the distribution specified by θ. However, the situation that we encounter in this thesis is when θ is *not* known; therefore we would like to give *some* values to these parameters. The Dirichlet distribution is a standard way to assign prior probabilities to the *k* multinomial parameters θ:

$$P(\theta) = \frac{1}{B(\alpha_1\ldots\alpha_k)}\prod_{i=1}^{k}\theta_i^{\alpha_i-1}$$

The parameters α of the Dirichlet distribution allow the introduction of prior knowledge about θ. Because it is a conjugate prior for the parameters of the multinomial distribution, the Dirichlet can easily be combined with the multinomial expression for the likelihood, according to Bayes' theorem (*cf.* Gelman, Carlin, Stern & Rubin, 1995) to specify the posterior distribution of θ:

$$P(\theta|c_1,\ldots,c_k) = \frac{1}{B(c_1+\alpha_1,\ldots,c_k+\alpha_k)}\cdot\frac{n!}{c_1!\ldots c_k!}\cdot\prod_{i=1}^{k}\theta_i^{c_i+\alpha_i-1}$$

[4] Strictly speaking, this view is restricted to the case where there is only one context window position. For larger window sizes, there may no longer be one unique outcome per 'trial' (*ie*. more than one valid context word may simultaneously occur in the window centered on *w*). Nevertheless, the likelihood of a co-occurrence vector can still be calculated using the same machinery.

For present purposes it is not even necessary to evaluate this equation, since we do not need to express the entire posterior density. Recall that our goal is to update the contextual distribution $P(\theta)$ (the semantic expectation) after encountering a context cue *c*; we can simply use one of the summary statistics for the Dirichlet posterior distribution to express this updated prior knowledge. For instance, we can update each prior parameter θ_i using the mean of the posterior distribution:

$$\hat{\theta}_i = \frac{c_i + \alpha_i}{\sum_{j=1}^{k}(c_j + \alpha_j)}$$

Here, the values of c_i are simply the component values of the vector representation for *c*. The effect of the Dirichlet parameters α (also known as the prior *weights*) on the posterior distribution can be seen as the observation of 'virtual' data. The main motivation for using this Bayesian update procedure is the straightforward way in which prior knowledge is combined with new data; with every iteration the posterior distribution serves as the prior distribution for the next cycle. By applying Bayes' rule iteratively to the words in an utterance, the prior distribution is revised on a word-by-word, incremental basis, and an updated semantic expectation is available at any point in time.

It is worth illustrating the computation of ICD for the *hermetically sealed* example. Imagine that we have constructed a three-dimensional semantic space, and that the corpus frequencies of the three context words $context_1$, $context_2$ and $context_3$ are 150, 120 and 50, respectively. Imagine also that we have extracted the co-occurrence vector representations (14, 1, 10) for *hermetically* and (55, 4, 41) for *sealed*. So, given the presence of *hermetically*, our task is to compute the difference between the expected meaning of the next word and its 'actual' meaning (*sealed*), using the relative entropy measure.

Recall that $P(\theta)$ represents the processor's expectations about the meaning of upcoming words. Because the shape of this distribution is initially unknown, we define $P(\theta)$ to be a Dirichlet distribution, and set the values of the Dirichlet parameters α_1, α_2, and α_3 to the corpus frequencies of the three context words (150, 120 and 50).

We next apply Bayes' theorem in order to update this inital prior with the co-occurrence data for *hermetically*, giving the Dirichlet posterior $P(\theta|hermetically)$. This updating simply involves revising each prior parameter θ_i using the above equation

for the posterior mean, resulting in the distribution (0.48, 0.35, 0.17). The final step is to calculate the ICD value for *sealed* as the relative entropy between the revised prior distribution (0.48, 0.35, 0.17) and the contextual distribution for *sealed* (0.55, 0.04, 0.41). The ICD value for this contrived example is 0.498 bits.

6.1.2 Weighting prior knowledge

In order to implement incrementality, we have given the ICD model two free parameters. These parameters are relevant to the shape of the posterior distribution, and thus also have potential impact on the model's psychological plausibility. The first concerns the issue of the *weighting* of prior information with new evidence. This question can be phrased simply as "how much weight should be given to prior knowledge?" Besides their interpretation as 'virtual' data, the parameters of the Dirichlet prior distribution can be viewed as weights. For example, if the sum of the prior weights (α_0) is 1,000, and the results of 100 new 'multinomial trials' are recorded, prior knowledge is effectively treated as 10 times more relevant than the new data. The next application of Bayes' rule will be to the updated prior distribution which reflects this 10:1 ratio.

The most straightforward weighting scheme involves holding the total prior weight (α_0) constant for each successive application of the update rule. Note that this procedure naturally implements a simple method of weighting according to *temporal proximity*; as more data are incorporated, earlier encountered data (*ie.* words occurring at the beginning of the utterance) have less and less influence on the shape of the posterior distribution.[5] The total prior weight was determined empirically, as detailed in section 6.1.3 below.

The second parameter of the ICD model is also concerned with weighting; the question posed is "should all new data be considered equal?" Recall that the new evidence consists of a co-occurrence vector that is interpreted as a set of *n* multinomial trials, where *n* roughly corresponds to the number of contexts in which the incoming word has occurred. Co-occurrence vectors for common words thus represent larger 'samples' of new data than co-occurrence vectors for rare words. For example, frequent words such as *look* will have more impact on the form of the updated prior distribution than less frequent words such as *gaze*. It is clear that very

[5] If α_0 is *not* held constant with each application of Bayes' rule, the order of words in the context has no impact on the 'final' prior distribution, and we have effectively implemented the 'bag-of-words' approach to representing context.

frequent words (*eg. off*) will 'swamp' the calculation of the prior – it is an empirical question whether this frequency sensitivity is a desirable feature or not. An alternative to this approach of implicitly 'frequency-weighting' new evidence is to hold the sample size constant, which means scaling the individual co-occurrence counts upwards or downwards until their total reaches a predetermined value. This is the approach taken here.[6]

6.1.3 Setting the total prior weight using predictive probabilities

How can the weighting of prior knowledge be determined? We set the value of α_0 empirically, by making a crucial, though uncontroversial assumption about the way natural language is generated. If minimising processing effort is an adaptive property of the human language processor, we would expect a certain degree of *semantic redundancy* to be present in natural language output. For efficient comprehension, it is advantageous if the meaning of one element of an utterance is informative about the meaning of the rest of the utterance. In other words, the process of understanding the intended message will be more efficient the more *predictable* it is. There is a further motivation for the presence of semantic redundancy which is identical to that posited for the incremental nature of interpretation: redundancy is attractive in order to maximise the chance of recovering partial meaning of an utterance if the input is clouded with noise.

Under the assumption that this principle of semantic redundancy is valid, we can roughly equate the redundancy present in an utterance with the average ability to predict the meaning of the next word (word n+1) from the current context. These predictions about meaning can be estimated using the mathematical machinery we have already introduced in this chapter. Conveniently, the denominator of Bayes' theorem describes the *posterior predictive distribution* for an observation, given prior knowledge:

$$P(c) = \int_{-\infty}^{\infty} P(c|\theta)P(\theta)d\theta$$

This equation expresses the predictive probability of a particular word vector c as the 'average' of the likelihoods $P(c|\theta_i)$, weighted by the prior probabilities $P(\theta_i)$.

[6] There are certainly other conceivable weighting schemes which may improve the cognitive plausibility of the model; however, we prefer to explore the simplest approach first and see how far that takes us.

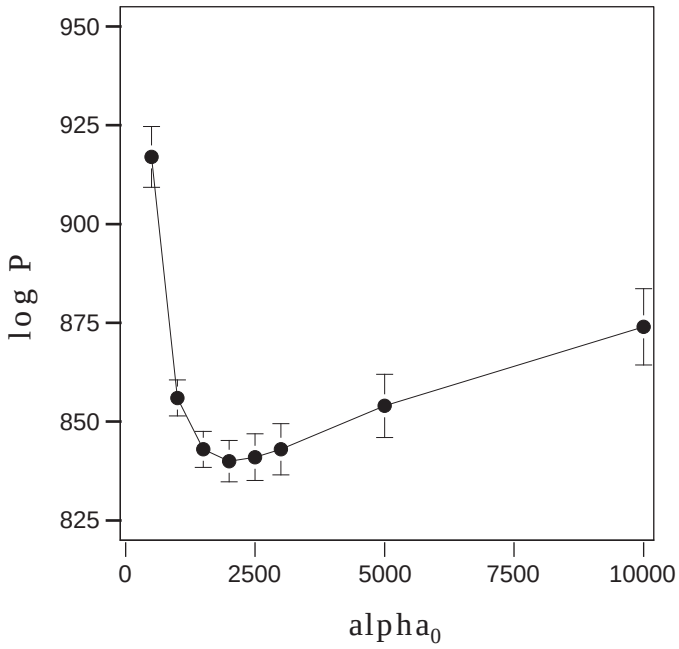

Figure 6-1. Mean predictive probability (with standard errors) of the word vectors in the text passage, as a function of α_0.

Computation of this quantity is trivial for the multinomial interpretation of a co-occurrence vector, because the predictive distribution of a Dirichlet posterior has a closed form:

$$P(c_1,\ldots,c_k) = \frac{n!}{c_1!\ldots c_k!} \cdot \frac{B(c_1+\alpha_1,\ldots,c_k+\alpha_k)}{B(\alpha_1,\ldots+\alpha_k)}$$

Thus, by calculating the predictive probabilities of the vector representations for each word in an utterance, the probability of the entire utterance can be estimated as the mean of the predictive probabilities for each word, and correspondingly the probability of an entire corpus can be estimated as the mean of the predictive probabilities for the words forming the corpus. We now have a simple empirical strategy for setting α_0: we manipulate this parameter in order to find the value which maximises the corpus probability.

Due to the computational expense of calculating predictive probabilities over a multi-million word corpus, we instead calculated this quantity for the words in a small sample of the BNC. This consisted of a passage of 94 sentences taken from a work of light fiction. From this passage, we only considered content words

(determined as in section 2.3.3), and then only those occurring at least 25 times in the BNC-spoken. This left 596 critical words.

Because the predictive probability is directly affected by the sample size (the number of 'multinomial trials'), it was necessary to set this to a constant value, by scaling the individual co-occurrence counts. The sample size was arbitrarily set to 1,000 'trials.' Since the predictive probability is the probability of the outcome of a set of multinomial trials – or alternatively, the probability of a high-dimensional vector representation – the space of possible outcomes is very large, meaning that the predictive probabilities for individual word vectors are extremely small. Therefore, probabilities were first logarithmically-transformed before plotting. Figure 6-1 displays the results of varying α_0 over a range of values (500 to 10,000). From the plot it appears that the corpus probability is maximised (–log *P* is minimised) when the value of α_0 is around 2,000. This is the value used in subsequent applications of the ICD model in this thesis.

6.1.4 Validation

We carried out a preliminary psychological validation of the ICD model by computationally simulating the same lexical priming experiment (Moss, Ostrin, Tyler & Marslen-Wilson, 1995, Experiment 2) analysed in Chapter 3. The variability in lexical processing effort induced by the presence of an immediately preceding prime word is ideal for testing the ICD model. Recall that we previously used this set of priming materials in order to demonstrate the psychological reality of a measure of 'semantic distance' between corpus-derived vector representations. In brief, the results of Chapter 3's simulation showed that Moss *et al.*'s semantic priming data could be explained in terms of representational similarity – word pairs that elicited human priming were more distributionally similar than matched unrelated pairs.

Our reanalysis of Moss *et al.*'s study using the ICD model was also successful. We tested the hypothesis that a minimal prior context – a single word – would have a reliable effect on the amount of information conveyed by the target word, and that this effect would pattern with the human behavioural data. Specifically, we predicted that a related prime word (such as *coat*) would *reduce* the amount of information conveyed by a target word (like *hat*) about its contexts of use, compared with an unrelated prime. The difference in ICD values resulting from the divergent influence of the related and unrelated prime words on the form of the

prior distribution was expected to correspond to the difference in processing effort (measured by lexical decision response times) observed by Moss *et al.*

We used the same 10M word text corpus to create co-occurrence vectors for the critical stimuli. Window size was five words before and after the target word, and the most frequent 500 content words in this corpus served as context words. The two ICD model parameters, the total prior weight (α_0) and sample size (n), were set to 2000 and 500, respectively.

The simulation results provided strong support for the plausibility of the ICD model. The overall pattern of semantic priming observed by Moss *et al*. (1995) was replicated by the difference in the ICD value for a target word (*hat*) presented after a related prime (*coat*), and ICD calculated for the same target word when presented after an unrelated prime (*pencil*). For example, ICD was determined to be 0.822 bits for *hat* when preceded by *coat*, and 1.190 bits for *hat* preceded by *pencil*. This difference was highly reliable across items, $F(1,92)=90.69$, $p<0.001$.[7]

This pilot test of the ICD model has opened up a novel way of understanding the effects of semantic context on word recognition. If the context (the prime) allows the processor to create precise expectations about upcoming meaning, processing of a word that conveys that meaning (a related target) is predicted to be facilitated. The simulation demonstrated that semantic priming can be interpreted as the influence of the local linguistic context on the quantity of information conveyed by a word about its contextual behaviour (*ie*. its meaning). The ICD model has successfully accounted for the variability in lexical processing effort observed using the standard semantic priming paradigm.

Our information-oriented explanation of the semantic priming effect stands in marked contrast to the spreading activation account predominant over the last 25 years. Spreading activation models (*eg*. Collins & Loftus, 1975) explain automatic priming in representational terms: encountering the prime word activates its conceptual (or semantic) entry in memory, and this activation is assumed to spread to other, semantically related representations. A target word that maps to one of these 'pre-activated' entries is therefore predicted to be recognised faster than if its conceptual representation is not primed. Thus, spreading activation theory provides

[7] The pattern of ICD differences was not completely identical to the pattern obtained using the Contextual Similarity measure (*cf*. section 0). Specifically, although there was again a Semantic Type × Relatedness interaction: $F(1,92)=6.97$, $p<0.05$, there was no evidence for the Association × Relatedness interaction reported by Moss *et al*. (1995): $F(1,92)<1$. This deficiency does not necessarily represent a principled failure of the ICD model, but merely that it is sub-optimal in some respect.

a *mechanistic* explanation of priming, by postulating an architecture where related concepts are linked, and an algorithm describing the flow of activation between concepts. In contrast, the ICD model is situated at a higher level; predictions about processing effort can be made without any assumptions about the actual cognitive mechanisms involved.

Based on a set of lexical priming experiments that failed to find differences in the size of the priming effect across a broad range of prime-target relationships, Hodgson (1991) argues that the spreading activation metaphor should be abandoned in favour of a semantic integration account; he considers priming to be the result of automatic post-lexical processes that are already needed for normal comprehension, in order to incorporate the meaning of a given word into the semantic representation formed from the preceding context. Although we do not wish to address the issue of the processing mechanism(s) responsible for priming, Hodgson's semantic integration theory appears to be compatible with the current information-oriented view. The ICD correlate of the single-word priming effect could be described as a reflection of the processing advantage resulting from the 'easier' integration of a target word's semantic representation with that of a related prime word, compared with an unrelated prime. However, in order to say something specific about the cognitive mechanism involved in lexical priming, one needs to carefully consider the large body of research (see, *eg*. Williams, 1996) concerned with this issue.

6.2 An information-oriented perspective on contextual constraint

In Chapter 4 we reviewed research on *contextual constraint*: the influence of the local linguistic context on the predictability of upcoming words. Recall that both Schwanenflugel and LaCount's (1988) Feature Description Model and our corpus-based Contextual Relevance Model define a constraining sentence context as a context that imposes detailed semantic feature restrictions on upcoming words, or alternatively, makes certain aspects of sentence meaning salient. Both models predict the attested effect that target words closely satisfying these meaning constraints are facilitated compared with targets that do not match as well.

Although we presented the Contextual Relevance Model as an improvement over the Feature Restriction Model, by virtue of its objectivity (it relies solely on the distributional information inherent in the linguistic environment), it has several

notable flaws (*cf.* section 4.4). To recap, one prominent limitation of this model is that it treats the context as a 'bag' of words, effectively considering syntactic dependencies between context words and word order to be irrelevant, which, in light of the behavioural evidence for incrementality, is clearly a false assumption. A more serious, related deficiency is its ignorance of constraints due to computational resources. The capacity of working memory certainly has an impact on the temporal processing involved in language comprehension and production (Caplan & Waters, 1999), and a parsimonious model should accommodate these constraints.

It is now appropriate to recast the *representational* approach to contextual constraint taken in Chapter 4 into the current information-oriented framework. The robust finding that words are processed more rapidly when preceded by a constraining context than by an unconstraining (or neutral) context can be explained in terms of the processor's expectations about the meaning of upcoming words. A highly constraining context allows the processor to create precise expectations about the meaning of upcoming words, and so the discrepancy between these expectations and the meaning of an appropriate upcoming word would be minimal, with a corresponding low level of processing effort. We propose that the ICD model provides a general, information-oriented explanation of the contextual constraint effect: a constraining context can be described as a particular context that *reduces* the quantity of information that is normally conveyed by a target word about its contexts of use. The constraint effect is modelled by the *difference* in the amount of information conveyed by a target word when preceded by a semantically constraining context, compared with a less constraining context.

Similarly, the ICD model also provides an explanation for the semantic congruity effect (*cf.* section 4.2). If the process of comprehension involves computing expectations about word meaning based on the linguistic context, then the semantic congruity effect simply reflects whether these semantic expectations are satisfied by the upcoming word. Processing is easier if the target word is semantically compatible with the context than if the target is incompatible.

Note that our new conceptualisation of the effect of contextual constraint is based on exactly the same functional principle put forward for the single-word lexical priming effect in section 6.1.4 above. Both phenomena are predicted from the expectation-building hypothesis, and both are captured by the same information-theoretic quantity. If the context – single word or sentence – is semantically related to the target word, then the amount of information conveyed by the target is

reduced. The ICD model permits a unified explanation of semantic context effects as the difference in the quantity of information conveyed about meaning.

It is apparent that the ICD model offers a potential improvement over the 'bag-of-words' approach to representing context. The amount of information conveyed by a sentence-final target word will depend on both the relative ordering of the words in the context and their temporal proximity to the target. If context words are scrambled, constraint effects diminish or disappear (*eg*. Masson, 1986), and the ICD model is predicted to behave accordingly.

Example (1) illustrates this property of the ICD model when applied to a highly constraining sentence fragment taken from Schwanenflugel's (1986) norms. Using a simple prior weighting scheme (settings: α_0=2000, n=500), *beach* in sentence (1a) provides 0.925 bits of information about its contexts of use, compared with 0.956 bits when calculated for (1b). As expected, *beach* conveys more information when following the scrambled sentence context, which predicts a slower response time for *beach* in (1b) compared with (1a). Of course, this is only one example; the prediction would need to be verified against more data.

(1) a.

	On	a	hot	summer	's	day	many	people	go	to	the	**beach**.
ICD	0.940		0.597	0.318			0.288	0.142				0.925
CD	0.940		0.608	0.318			0.199	0.131				0.991
difference	0.000		0.011	0.000			-0.089	-0.011				0.066

b.

	Many	to	summer	's	on	go	hot	the	a	people	day	**beach**.
ICD	0.608			0.142		0.885				0.311	0.325	0.956
CD	0.608			0.131		0.940				0.199	0.318	0.991
difference	0.000			-0.011		0.055				-0.112	-0.007	0.035

A useful feature of the ICD model is that predictions of processing difficulty are available at *every* point in the context. Contextual constraint is assumed to be an incremental phenomenon; as more of the words comprising a highly constraining sentence context become available to the comprehender, constraint strength should tend to increase, reducing the effort of processing each forthcoming word. However, ICD values would not generally decrease as Bayes' rule is applied to successive words an utterance, because ICD values also reflect the 'base' contextual distinctiveness value of the word itself. We might instead expect the *difference* between context-independent CD and ICD to monotonically increase when moving forward through the context, if a semantically constraining context systematically reduces the amount of information conveyed by subsequent words. This does not appear to be true, at least for this example.

6.2.1 Eye movement data

One useful tactic for studying the influence of contextual constraint on lexical processing is through analysis of the eye movements that people make while reading silently. Eye-tracking technology allows an accurate temporal record to be made of the on-line processing of natural language, and analysis of several types of eye movement measurements can give some insight into the dynamic processing that is involved in normal reading (*eg.* Rayner, Sereno, Morris, Schmauder & Clifton, 1989). Recording eye movements seems ideal for testing word-by-word predictions about the effects of contextual constraint on natural reading processes.

Dependent variables such as gaze duration (the total length of time the eyes fixate a word before leaving it), first fixation duration (the duration of the initial fixation made on a word), probability of fixating/skipping, and number of regressions (the number of times the eyes return to a word from a point forward in the text) can be measured on the same word type embedded in different linguistic contexts. However, these measures obviously cannot *isolate* the effects of context from other influential determinants of eye movement behaviour; for instance, perceptual properties such as word length, and textual properties such as line position and occurrence at a sentence boundary all affect fixation probability and duration (*eg*. Kliegl, Olson & Davidson, 1982). Lexical variables such as corpus frequency have also proved to be important (Rayner & Duffy, 1986). Reichle, Pollatsek, Fisher and Rayner (1998) describe the relationship between lexical properties and processing behaviour as follows:

> … [L]exical and/or semantic characteristics of a word – or something closely related to them – appear to be able to control the duration of the fixation on that word, and, thus, the relation between cognitive processes and eye movement control is fairly tight (p. 127).

Eye movement measures also cannot easily pin down the *locus* of context effects on lexical processing; for instance, fixation times seem to reflect the processing required to access or identify a target word, as well as the effort involved in integrating it into the current discourse context (Balota, Pollatsek & Rayner, 1985; Rayner *et al*., 1989). Since the ICD model is presented as a computational-level explanation, we shall not attempt to determine the particular stage of lexical processing (*ie*. access/identification or integration into the discourse-level semantic representation) that is affected by context.

The empirical finding most relevant for this thesis is that contextual constraint has a robust effect on eye movement measurements made during reading:

contextually constrained words are fixated for less time and are skipped more often than words that are not constrained by the semantic context (*eg*. Altarriba, Kroll, Sholl & Rayner, 1996; Rayner & Well, 1996). This result is consistent with semantic constraint effects observed using the visual lexical decision task (*eg*. Schwanenflugel & Shoben, 1985; *cf*. section 4.2).

The change in constraint strength resulting from varying even one word in the immediate context of the target has been shown to influence eye movement behaviour. For example, Schustack, Ehrlich and Rayner (1987) found that the probability of fixating on a target noun (*eg*. *floor*) was significantly lower when it was preceded by a 'semantically restrictive' (*ie*. constraining) verb (*sweep*) compared with a less restrictive verb (*clean*), in the same paragraph context.

These findings mesh easily with the predictions of the ICD model. If a restrictive verb allows the processor to form more precise expectations about the meaning of its direct object than a less restrictive verb, then less effort should be required to process a suitable direct object. The amount of information conveyed by a compatible target noun should vary according to the restrictiveness of the verb. Generally speaking, the ICD model should be sensitive to the fact that a verb such as *sweep* constrains the realisation of its direct object to a greater degree than a verb like *clean*. Experiment 14 tested this hypothesis through reanalysis of Schustack *et al*. (1987, Experiment 2).

6.2.2 Experiment 14[8]

In this experiment, we gave the information-oriented definition of contextual constraint a preliminary evaluation, by submitting Schustack *et al*.'s (1987) stimuli and their computed ICD values to a by-item analysis, treating ICD as the dependent variable. If the ICD model successfully captures the constraint variable that affects eye movement measures, then we would expect an ICD effect for the same target word presented in paragraph contexts with varying amounts of contextual constraint. Schustack *et al*. (1987, Experiment 2) manipulated constraint by varying the 'semantic restrictiveness' of the verbs immediately preceding (and syntactically associated with) a set of 40 target nouns. Their Constraint factor had two levels, General and Restrictive; verbs were generated for each condition according to experimenter intuitions. A norming study indicated that the Restrictive verbs were

[8] Experiments 14 and 15 were originally presented in McDonald (1999).

significantly more predictive of the target noun than were the General verbs. An example of a critical sentence from their materials is given in (2); half of the subjects saw *hung*, the Restrictive, or constraining context, and half saw *put*, the General, or less semantically constraining verb.[9]

(2) He *hung*/*put* the **picture** on the wall that had the biggest crack.

Constraint had a significant effect on three different eye movement measures: fixation probability, gaze duration and total fixation time. Target nouns in sentences containing a Restrictive verb had a lower probability of being fixated, and were fixated for a shorter time than targets presented after the General verbs. As Schustack *et al.* mention, these eye movement data cannot isolate the locus of the constraint effect; that is, whether semantic restrictiveness influences word recognition or integration of the target word into the semantic representation of the discourse.

There are many potential variables which might underlie the constraint effect reported by Schustack *et al.*; the goal of the present reanalysis was to determine if the effect could be accounted for by the ICD value of the target noun when preceded by the minimal verbal context, as predicted by the model of incremental processing introduced in this chapter.

6.2.2.1 Method

Inflected word forms in the material set were first converted to their canonical (lexeme) forms. From the 40 original items listed in Schustack *et al.*, six were removed because either the Restrictive verb or the target noun had a BNC-spoken lexeme frequency of less than 25 occurrences. The mean log lexeme frequency of the General verbs was 8.059, compared with 5.948 for the Restrictive verbs.

Co-occurrence statistics from the BNC-spoken provided vector representations for the critical stimuli. Window size was ±5 words, with the most frequent 500 content words used as context words. We calculated ICD values for each target noun in both conditions, using the Bayesian update procedure described earlier, with the total prior weight (α_0) set to 2,000, and sample size (n) set to 500 "trials".

[9] Schustack *et al.* also manipulated a second independent variable, the recency of prior mention of the target word, which did not interact with Constraint.

6.2.2.2 Results and Discussion

For 24 out of the 34 items, the ICD value computed for the target noun in the Restrictive condition was smaller than its value in the General condition. The mean quantity of information conveyed by a target word following a General verb was 1.084 bits, compared to 1.031 bits after a Restrictive verb. Because the differences within pairs were not normally distributed, the Wilcoxon signed-ranks test was used to test the null hypothesis that the difference in ICD values between the two conditions was due to chance. This difference was statistically significant: z=2.530, p<0.01, one-tailed. Target nouns such as **picture** conveyed less information about their contexts of occurrence when preceded by a Restrictive predicate like *hung* than when preceded by a General verb such as *put*. The effect of the semantic constraint imposed by the Restrictive verbal context has been to *reduce* the quantity of information conveyed by the target noun (when compared with the General verbal context). The ICD model has captured the effect of the highly constraining local context on eye movement measurements.

In terms of the expectation-building strategy proposed in Chapter 1, we could say that Restrictive verbs allow the formation of more precise expectations about the meaning of their direct objects than General verbs; the effect on eye movement behaviour simply reflects the better fit between the expectations formed by a Restrictive predicate with the meaning evoked by its direct object.

Although Schustack *et al*. invoke the spreading activation metaphor in order to characterise the pattern of facilitation they found, other researchers have attempted to explain the effect of contextual constraint on eye movement variables by adopting the Feature Restriction Model (*eg*. Altarriba *et al*., 1996). The results of our reanalysis of Schustack *et al*. (1987) indicate that the ICD model is able to account for the effects of contextual constraint using an entirely different approach. Exploiting the purely objective source of information available from simple distributional statistics appears promising for explaining the context-dependent variation observed in human reading performance.

6.2.3 Experiment 15

Although indicative, Experiment 14 was limited in that it did not test the predictions of the ICD model when confronted with a context longer than one word. The contextual constraint effect from the content word occurring immediately before the target was demonstrated to be capturable by ICD; it remains to be seen whether

this relatively simple model is still effective with larger amounts of context. Therefore, the aim of Experiment 15 was to further evaluate the ICD model, using pairs of multi-word sentence contexts that varied in how strongly they constrained the same target word. The materials from Altarriba *et al.*'s (1996, Experiment 1) eye movement study were ideally suited; these items produced reliable effects of constraint on skip probability and first fixation duration. (Contextual constraint was determined using a cloze procedure). Altarriba *et al.* also manipulated the target word frequency, finding evidence for a frequency effect on gaze duration, but no interaction between the two variables. In this by-item reanalysis, we also examined the influences of both the Constraint and Frequency factors, treating ICD as the dependent variable. We predicted independent effects of each factor on ICD. A Frequency effect was anticipated because of the natural correlation between CD and frequency (*cf.* section 5.2.2): in general, high frequency words tend to convey less information about their contexts of use than low frequency words.

6.2.3.1 Method

Preparation of the materials was virtually identical to Experiment 14. The original stimuli consisted of 32 high-frequency and 32 low-frequency target words, with a high- and low-constraint sentence context created for each target. Four of the low-frequency items were removed; their critical target words failed to meet the lexeme frequency threshold of 25 occurrences in the BNC-spoken. The high-frequency targets had a mean log lexeme frequency of 7.002, compared with 4.736 for the low-frequency targets. The sentence contexts for the high-frequency target word **teeth** are displayed in (3), and the contexts for the low-frequency target **thief** are given in (4):

(3) a. The dentist told me to brush my **teeth** after every meal. (*high-constraint*)
b. He lost three **teeth** and had a black eye after the fight. (*low-constraint*)

(4) a. The robbery was committed by a **thief** who was known for his skill in safe-cracking. (*high-constraint*)
b. He warned us that the **thief** had escaped from prison on Wednesday. (*low-constraint*)

Targets and content words in the context preceding each target were first replaced with their canonical forms, if necessary, and co-occurrence vectors were created using the same procedure as in Experiment 14. ICD values for each target word were computed using Experiment 14's method and parameter settings.

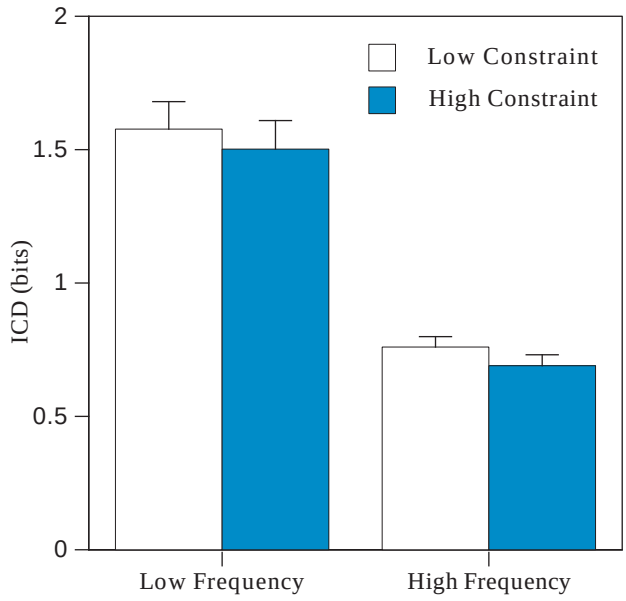

Figure 6-2. Computed ICD values as a function of contextual constraint and word frequency (with standard errors).

6.2.3.2 Results and Discussion

Figure 6-2 graphically displays the results of the simulation. The pattern of ICD values closely replicated the pattern of eye movement measurements reported by Altarriba *et al*. ICD values were smaller for high-constraint contexts than for low-constraint contexts, and were also smaller for high-frequency target words compared with low-frequency targets. An analysis of variance indicated reliable main effects of both Constraint and Frequency: $F(1,58)=22.15$, $p<0.001$, and $F(1,58)=58.53$, $p<0.001$, respectively. There was no interaction: $F(1,58)<1$. Separate ANOVAs on the high-frequency and low-frequency items also revealed significant Constraint effects: $F(1,31)=17.44$, $p<0.001$, and $F(1,27)=7.83$, $p<0.001$, respectively.

The results of Experiment 15 constitute further empirical support for the ability of the ICD model to capture the effects of contextual constraint on lexical processing behaviour. Building on the successful simulation of the effect on eye movements of varying a single word in the preceding linguistic context (Experiment 14 above), ICD has also been shown to account for observed constraint effects when reading a word in a two different, yet equally plausible sentential contexts. The simulation results corresponded closely to Altarriba *et al.*'s eye movement data. The

amount of information conveyed by a target word was reliably smaller when the word was presented in a highly constraining context compared with a less constraining context. Because the independent effect of target word frequency was also replicated, it is clear that ICD does not simply measure contextual constraint; it also incorporates lexical (or semantic) factors that affect eye movement behaviour. The ICD model has simultaneously captured variability in lexical processing effort attributed to both the frequency of the target word and its predictability in context.

6.2.4 Experiment 16: Analysis of an eye-tracking corpus

Recording eye movements during silent reading is one of the least manipulative procedures for investigating the effect of a constraining prior context on the processing of upcoming words – it is certainly more ecologically valid than traditional word recognition tasks such as lexical decision or naming, since overt responses are not required. However, experiments that examine context effects using eye movement methodology (including the two studies submitted to computational reanalysis above) typically use contrived sentence materials. Stimulus sentences are usually constructed to be homogenous (of similar length and syntactic complexity), and are presented one at a time, which is obviously different from how passages of text are normally encountered. It is not entirely clear that results obtained under these conditions fully generalise to normal reading processes.

It is possible, however, to investigate reading behaviour using more ecologically valid methods. By having a group of subjects read selections of natural text, a corpus of eye movement data can be collected and further analysed to test any number of hypotheses. Moreover, the corpus analysis approach offers an opportunity to test the predictions of comprehensive models of reading that attempt to account for word-by-word processing variability (*eg*. Just & Carpenter, 1980). In Experiment 16, we apply this corpus strategy in order to test the ICD model's predictions about lexical processing effort, on a word-by-word basis.

Although there is compelling evidence that the amount of time a reader's eyes spend on a word is a transparent indicator of lexical processing difficulty, the relationship between fixation time and processing effort is certainly not perfect. Rayner *et al*. (1989) point out that (a) processing of a target word is sometimes initiated on the *previous* fixation (because of parafoveal preview), and (b) processing can 'spillover' onto the next word: if word *n* is difficult to process, fixation time on word *n*+1 increases. Both gaze and first fixation duration are

```
His head was full of sentences he was going to write to Hilary
when he had the time to put pen to paper: I may remind you that I
never asked you for a penny towards the summer gas bill ... do
you think I am made of stone?  ... surely I deserve better
consideration ... who listened for hours when you had that
```

Figure 6-3. A fragment of the 1,000 word text passage used to create the eye movement corpus.

problematic as measures of per-word processing time for these reasons. In addition, words can be processed even if they are *not* fixated; this is particularly true of functors and other short words. Consequently, the probability of fixation/skipping may reflect a word's perceptual properties to a greater extent than the lexical (or semantic) properties that determine the ease of processing. Since skipping word *n* generally inflates the fixation time on word *n*–1, fixation duration is also sensitive to perceptual factors. Rayner (1998, p. 377) argues that "... any single measure of processing time per word is a pale reflection of the reality of cognitive processing." Following Rayner's recommendation, we examined several eye movement measures, adopting the view that more substantiated inferences about on-line comprehension can be drawn from the eye movement record through analysis of more than one dependent variable.

The statistical approach we used to investigate the relationship between ICD values and the various eye movement variables is multiple regression analysis. Because the variability in processing effort due to the ongoing influence of the semantic context is primarily of interest, it is necessary to control for the influence of extraneous variables – textual and perceptual factors – which can be done using regression techniques. The aim of the regression analyses reported below was to quantify the amount of reading variability accounted for by ICD, once the effects of perceptual and text-structure variables had been partialled out. Our approach was to remove the variance attributable to orthographic word length, line position (beginning and end), and occurrence at the end of a clause – all influential factors that tend to lengthen fixation durations (*eg.* Just & Carpenter, 1980; Kliegl, Olson & Davidson, 1982) – permitting estimates of the amount of eye movement variance uniquely explained by ICD.

6.2.4.1 Method

An eye-tracking corpus was available for the proposed analyses.[10] This corpus consisted of eye movement records for ten subjects, each reading the same passage of 46 sentences (approximately 1,000 words) taken from Beryl Bainbridge's novel *An Awfully Big Adventure*, which forms part of the BNC. A sample fragment of the corpus is displayed in Figure 6-3.

The corpus analysis consisted of two parts. In Part 1, we looked at the relationship between the context-independent CD measure and three eye movement variables, and in the second part we investigated the relationship between context-dependent ICD and the same dependent variables.

Because CD was shown to be a significant predictor of the processing difficulty reflected by lexical decision latencies (*cf.* section 5.2.1), it is worthwhile to further assess its predictive power against measures of reading performance. For each word of interest in the text passage, we measured first fixation duration (the length of the first fixation made on the target word), gaze duration (the total time spent on a word prior to a saccade to another word), and the skip probability (the proportion of subjects who failed to fixate upon the target word at all). Although gaze and first fixation duration are typically highly correlated, it is advantageous to examine both measures in case a distinction is discovered.

We also assessed the predictive power of two highly influential lexical variables: word length and frequency. Length in letters has been shown to strongly influence both fixation times and the probability of skipping; gaze duration and fixation probability increase with word length (*eg.* Rayner & Duffy, 1986). Corpus frequency is also a robust predictor (*eg.* Altarriba *et al.*, 1996; Just & Carpenter, 1980; Rayner & Duffy, 1986): fixations made on high-frequency words are generally shorter than fixations on low-frequency words. Our hypothesis is that while word length and frequency should account for significant amounts of gaze duration and skip probability variance, CD will not, even after lexical and text-structure factors are partialled out. We have claimed earlier (Chapter 5) that the CD measure captures aspects of the effort involved in processing *isolated* words; thus we expected the

[10] Eye movement recordings were made by Padraic Monaghan and Louise Kelly of the University of Edinburgh, using a Dual Purkinje Eyetracker belonging to the Department of Psychology, University of Glasgow.

Table 6-1. Correlation Matrix for Six Predictor Variables (*n*=381).

Variable	1	2	3	4	5	6
1. LineB	1.00					
2. LineE	-0.10	1.00				
3. ClauseE	0.01	0.13**	1.00			
4. WL	0.13*	0.05	0.01	1.00		
5. lnLF	-0.04	-0.11*	-0.08	-0.42**	1.00	
6. CD	0.03	0.11*	0.08	0.35**	-0.82**	1.00

Note: Correlation coefficients involving dichotomous variables are point biserial *r*'s and ϕ coefficients. *p<0.05 **p<0.01 (2-tailed).

ICD model to provide a better account than CD of context-dependent processing behaviour, since ICD was specifically designed to take context into consideration.[11]

In Part 2 of the analysis, we evaluated the predictions of the ICD model against the same three dependent variables. Specifically, we anticipated a linear relation between ICD and the reading measures, once extraneous 'non-semantic' variables such as word length and line position were partialled out. We assume that some portion of the remaining eye movement variance is attributable to the constraining effect of the preceding semantic context, and that this variability can be explained as the amount of information conveyed by the target word. We expected ICD values to mirror constraint effects due to both predicate 'restrictiveness' (as demonstrated in Experiment 14) and the existence of semantic relationships between the target and words in its immediate sentential context (*eg*. Morris, 1994). The hypothesis tested was whether a significant portion of the word-by-word influence of context on eye movements during reading could be explained by the ICD model.

6.2.4.2 Analysis Part 1

In the first part of the eye-tracking corpus analysis, we examined the relationship between a word's contextual distinctiveness (CD) score and the three dependent variables: gaze duration (Gaze), first fixation duration (FirstF), and the probability of skipping (SkipP). Since word length (WL) and corpus frequency (lnLF) are typically highly influential factors, they were included in the multiple regression analyses together with three dichotomous textual variables: occurrence in line

[11] It is intuitively plausible that providing contextual support for very high-CD words like *amok* should substantially reduce processing effort, due to the construction of semantic expectations. The ICD model predicts that the processing of *amok* immediately preceded by *run* should be easier than the processing of *amok* presented in isolation.

Table 6-2. Summary of Individual Multiple Regressions on Gaze Duration.

	LineB	LineE	ClauseE	WL	lnLF	CD
Mean *B*	8.99	36.14	37.54	6.78	-5.24	0.19
SE	12.89	18.64	16.95	1.45	1.90	5.75
t(9)	0.70	1.94	2.21	4.68	-2.76	0.03
1-tailed *p*	0.252	0.042	0.027	0.001	0.011	0.487

beginning position (LineB), line end position (LineE) and clause-final position (ClauseE).[12] We eliminated fixations of less than 100 ms from further analysis (for justification, see Rayner *et al.*, 1989). Note that the analyses were carried out for the 381 content word tokens in the text passage that had a BNC-spoken lexeme frequency of 25 or more (to ensure reliable CD scores). The zero-order correlations between the six independent variables are shown in Table 6-1, and the descriptive statistics for Gaze, FirstF and the three non-dichotomous predictor variables are displayed in Table 6-5. Only WL, CD and lnLF had any appreciable relationship with each other; these correlations were in the directions anticipated from the results of the word recognition studies in Chapter 5. Unfortunately, the high intercorrelation between lnLF and CD (r=-0.82) means that collinearity will be a problem for interpretation of the regression analyses.

Gaze duration

In accordance with Lorch and Myers' (1990) recommendations, regression equations were separately fitted to each subject's Gaze data (only words that were actually fixated were included), and one-sample *t* tests were used to assess the reliability of the six partial regression coefficients. The results of the individual regression analyses are displayed in Table 6-2. There were reliable unique effects of both LineE and ClauseE, indicating that gaze durations on words occurring at the end of a line or at the end of a sentence tended to be longer than for words occurring in other positions. WL was a significant predictor of Gaze; subjects spent more time fixating long words compared with short words. There was also a reliable independent

[12] Conducting by-item analyses entails collapsing over subjects. Multiple regression analyses carried out on these averaged fixation times would not take between-subject variance into account, which means that although the partial regression coefficients would be unbiased estimators of the population values, the results of testing each coefficient for significance could not be generalised beyond the sample (for further discussion, see Lorch & Myers, 1990).

Table 6-3. Summary of Individual Multiple Regressions on First Fixation Duration.

	LineB	LineE	ClauseE	WL	lnLF	CD
Mean *B*	8.90	0.88	4.41	-0.72	-3.95	1.15
SE	4.66	7.18	4.49	0.93	1.01	3.23
t(9)	1.91	0.12	0.98	-0.77	-3.92	0.35
1-tailed *p*	0.045	0.453	0.176	0.231	0.002	0.366

influence of lnLF on Gaze; gaze durations increased as word frequency decreased. CD, as anticipated, was not a significant independent predictor of Gaze when variance attributable to the other factors was partialled out.

First fixation duration

Individual regression analyses were conducted for the ten participants' FirstF data (see Table 6-3). After partialling out the variance shared with the other variables, there were reliable relationships between first fixation duration and LineB and lnLF only. The first fixation made to a word was significantly longer if it occurred at the beginning of a line than in other positions. The significant effect of lnLF on FirstF parallelled the results of the gaze duration analysis; as word frequency increases, first fixation duration tends to decrease. There was again no unique role for CD.

Skip probability

First, the zero-order correlations between the probability that a word was skipped during reading (SkipP) and the five predictors were computed. WL, lnLF and CD were all significantly correlated with SkipP: r=-0.64, p<0.001, one-tailed; r=0.38, p<0.001; r=-0.26, p<0.001, respectively. However, this represents a by-item analysis, since the probability of skipping a word is defined as the proportion of subjects who failed to fixate on the word. We next conducted separate logistic regression analyses for each participant, in order to determine which factors were unique predictors of whether a word was skipped or not (a dichotomous variable).

Table 6-4. Summary of Individual Multiple Regressions on Skip Probability.

	LineB	LineE	ClauseE	WL	lnLF	CD
Mean *B*	-0.62	-0.54	-0.11	-0.52	0.12	0.22
SE	0.73	0.33	0.21	0.04	-0.05	0.15
t(9)	-0.85	-1.65	-0.05	-12.10	2.51	1.46
1-tailed *p*	0.210	0.067	0.480	0.000	0.017	0.089

Table 6-5. Descriptive Statistics for Gaze and First Fixation Duration and Four Predictors.

	Gaze	FirstF	WL	lnLF	CD	ICD
n	3018	3018	381	381	381	381
Min	100	100	2	3.256	0.050	0.102
Max	1661	1532	13	11.344	3.294	3.319
Mean	305	265	5.6	7.042	0.848	0.871
Std Dev	147	105	2.0	1.947	0.599	0.596

Results of these analyses are displayed in Table 6-4. There were significant unique effects of WL and lnLF only. Long words were less likely to be skipped than short words, and frequent words were more likely to be skipped than rare words; these findings corroborate previous research (*eg*. Rayner & Duffy, 1986). Despite the significant negative correlation obtained between SkipP and CD in the by-item analysis, there was no reliable unique effect of CD when between-subjects variance was taken into account and the other variables were included in the regression.

In summary, the first part of the eye-tracking corpus analysis has generally confirmed our predictions regarding the relationships between the three eye movement variables and WL, lnLF and CD. A reliable WL effect was observed for Gaze and SkipP, but not for FirstF. Long words were less likely to be skipped and tended to be fixated for a greater amount of time than short words. Significant unique roles for lnLF were found for all three dependent variables: common words were more likely to be skipped than rare words, and of the words that were fixated, frequent words were fixated for less time than rare words. Two textual variables, LineE and ClauseE, had independent influences on Gaze, which is consistent with findings of previous eye movement research (*eg*. Kliegl, Olson & Davidson, 1982). The prediction that CD would fail to explain a unique portion of eye movement variance was also confirmed.

6.2.4.3 Analysis Part 2

In the first part of the eye-tracking corpus analysis, we did not find any evidence for CD as a reliable unique predictor of eye movement behaviour; the amount of information provided by a word about its contexts of use did not appear to be related to measures of context-dependent reading performance. This finding is consistent with the context-dependent nature of lexical processing, since a number of studies have shown that the effort of processing a given word depends on the

Table 6-6. Summary of Individual Regressions on Gaze Duration.

	LineB	LineE	ClauseE	WL	lnLF	ICD
Mean *B*	7.97	39.53	38.15	6.71	-3.69	3.18
SE	12.99	17.70	17.28	1.54	1.67	4.59
t(9)	0.61	2.23	2.21	4.37	-2.21	0.69
1-tailed *p*	0.278	0.026	0.028	0.001	0.028	0.253

Table 6-7. Summary of Individual Regressions on First Fixation Duration.

	LineB	LineE	ClauseE	WL	lnLF	ICD
Mean *B*	9.25	3.49	5.06	0.23	-3.34	1.43
SE	4.00	6.31	4.75	0.93	1.05	3.45
t(9)	2.31	0.55	1.06	0.25	-3.18	0.41
1-tailed *p*	0.023	0.297	0.158	0.405	0.006	0.345

Table 6-8. Summary of Individual Regressions on Skip Probability.

	LineB	LineE	ClauseE	WL	lnLF	ICD
Mean *B*	-0.63	-0.38	0.04	-0.50	0.10	0.11
SE	0.73	0.31	0.22	0.04	0.05	0.16
t(9)	-0.85	-1.23	0.19	-11.64	2.05	0.69
1-tailed *p*	0.209	0.126	0.427	0.000	0.035	0.253

linguistic context, and the CD measure does not take context into account. It is clear that the prior context provides a conditioning environment for interpretation of a particular word during reading; the effort of recovering the meaning of that word is influenced by what has already been seen.

As stated earlier, a *constraining* prior context intuitively provides cues about the meaning of an upcoming target word (increasing its predictability, in some sense), which Schustack *et al.* (1987) and Altarriba *et al.* (1996) have shown to facilitate the processes involved in reading – identifying a word and/or integrating it into the previous discourse. Conversely, a *neutral* context is less informative about the meaning of upcoming words, which corresponds to a greater degree of processing effort. Thus, the basic behavioural prediction is that the more informative the preceding context is about the meaning of an upcoming word, the less processing effort incurred when reading that word – realised behaviourally as a reduction in fixation time and a greater chance of being skipped. If, as claimed earlier, the ICD model quantifies the difficulty of processing a word *in context*, ICD should succeed

in explaining a unique amount of the variance in the eye movements made during reading.

The same independent variables examined in Part 1 (with ICD replacing CD) were assessed against Gaze, FirstF and SkipP, again by calculating regression equations separately for each subject. ICD values were computed with the total prior weight (α_0) and sample size (n) parameters set to 2,000 and 500, respectively, and only content words occurring in the five word window before the target contributed to the computation. The size of this window was chosen to roughly approximate the extent to which previous context could influence the processing of the target word. Note that this window, like the context window used when extracting co-occurrence statistics, ignored sentence boundaries.

The results were very similar to those of Part 1. In the Gaze analysis, significant independent effects were found for LineE, ClauseE, WL and lnLF only (see Table 6-6). Significant unique amounts of FirstF variance were accounted for by LineB and lnLF (Table 6-7). Only WL and lnLF accounted for unique amounts of variance in the logistic regression analyses conducted on SkipP (Table 6-8). ICD was not a reliable unique predictor of eye movement behaviour for any of the measures.

To summarise, Part 2 of the corpus analysis failed to support ICD as an independent predictor of the eye movement behaviour recorded during the reading of a text passage. Part of the difficulty in detecting such an effect may be due to the nature of the text passage itself. Because it was not explicitly manipulated, contextual constraint simply might not have varied enough to have a measurable impact on the eye movement record. For instance, Hyona (1993) did not find an anticipated constraint effect on reading behaviour when using materials with a somewhat restricted range of constraint (his high-constraint target words had a mean cloze value of 65%, compared with 32% for his low-constraint targets).

A second obstacle preventing discovery of a unique effect of ICD is due to collinearity; frequency and ICD were highly intercorrelated and it is difficult to separate their contributions using multiple regression analysis. The predictive power of ICD is clarified when individual multiple regression analyses are conducted *without* including frequency. When lnLF is not in the equation, ICD accounts for significant unique amounts of Gaze (Table 6-9), FirstF (Table 6-10) duration variance, and a marginally significant amount (p=0.056) of SkipP variance (Table 6-11). Experiment 15 showed that both constraint and frequency effects could be simultaneously captured by the ICD model; in terms of the current regression

Table 6-9. Summary of Individual Regressions on Gaze Duration (without lnLF).

	LineB	LineE	ClauseE	WL	ICD
Mean *B*	8.15	39.63	38.52	7.15	12.06
SE	13.00	17.72	17.42	1.44	2.04
t(9)	0.63	2.24	2.21	4.96	5.90
1-tailed *p*	0.273	0.026	0.027	0.001	0.000

Table 6-10. Summary of Individual Regressions on First Fixation Duration (without lnLF).

	LineB	LineE	ClauseE	WL	ICD
Mean *B*	9.22	3.70	5.44	0.65	9.37
SE	4.00	6.35	4.79	0.96	2.43
t(9)	2.30	0.58	1.13	0.68	3.85
1-tailed *p*	0.024	0.254	0.143	0.257	0.002

Table 6-11. Summary of Individual Regressions on Skip Probability (without lnLF).

	LineB	LineE	ClauseE	WL	ICD
Mean *B*	-0.61	-0.40	0.03	-0.51	-0.20
SE	0.73	0.31	0.22	0.04	0.11
t(9)	-0.83	-1.31	0.12	-11.53	-1.77
1-tailed *p*	0.213	0.112	0.454	0.000	0.056

analysis, it is preferable to include a single, general variable if one is available, than retain two separate variables which end up competing for variance. On these grounds, ICD is the more theoretically attractive variable.

6.2.5 General discussion

In section 6.2 above, we advanced an information-oriented reconceptualisation of contextual constraint, and we suggested that the ubiquitous single-word priming effect could be viewed as one particular manifestation of a more general effect: the constraints on the meaning of upcoming words imposed by the preceding linguistic context. The results of our analyses of eye movement data constitute further support for a unified explanation of semantic context effects. The definition of contextual constraint should be expanded beyond the notion of predictability in context, in order to include a wide range of contextual influences that affect

measurements of lexical processing effort. We have shown that the ICD model accommodates these influences under a single information-theoretic framework. Common to each of our computational reanalyses of constraint effects on the processing of a target word – whether attributed to semantic priming, 'semantic restrictiveness', or predictability – is the *reduction* in information conveyed about its meaning.[13] The attraction of the ICD model is that it provides a unified explanation of a range of semantic context effects, and this explanation relies only on the distributional information inherent in the linguistic environment.

Experiments 15 and 16 have also shown that in addition to capturing semantic influences on eye movements, ICD incorporates information about a word's frequency of use. ICD is naturally correlated with word frequency, and thus predicts the attested influence that frequency has on lexical processing effort. ICD is a theoretically more interesting variable than frequency, because unlike frequency, it also incorporates the effects of contextual constraint.

One additional way to view ICD values is in terms of Chapter 1's discussion of processing efficiency. The ICD model provides a quantitative estimate of the effect that prior knowledge (the semantic information present in the preceding context) has on the recovery of target word meaning. In other words, ICD supplies an estimate of the relative benefit or cost to lexical processing that can be attributed solely to the linguistic context in which the target word appears. This feature allows any number of potential contexts to be evaluated *a priori* for their relative ability to facilitate or impede the processing of a subsequent target word.

Finally, the most appealing feature of the ICD model is that it provides an integrated account of both context-*independent* and context-*dependent* processing phenomena; the effort of processing a word in isolation *or* in context is estimated by the quantity of information the word provides about its contextual behaviour. In both cases, processing effort is construed as the difference between the processor's expectations about a word's meaning and its 'actual' meaning; the presence of a semantically constraining context results in the formation of more precise

[13] A direction worth exploring in future is the relationship between ICD and *spillover* effects on the duration of eye fixations. It may be the case that spillover effects can also be parsimoniously explained using the ICD model. Rayner and Duffy (1986) found that the time spent fixating a target word was increased if it was preceded by a low-frequency word. For example, processing of the verb *moved* was easier when preceded by *vehicle* than by low-frequency *gondola*. This is the same prediction the ICD model would make, but for different reasons: the ICD value of *moved* occuring after *vehicle* would be lower than after *gondola* because *vehicle* permits the processor to form more precise expectations about the meaning of the next word.

expectations. No assumptions about the actual cognitive mechanism(s) involved that implement the expectation-building strategy are required to model a diverse set of empirical data.

6.3 Postscript: the prediction of upcoming words

This section represents a moderate diversion from the preceding material, in order to discuss some interesting implications raised by the ICD model. Could the model be applied to the task of *predicting* upcoming words? Word prediction is an important task for many statistical NLP applications; could the ICD model make a contribution here?

6.3.1 *N*-gram language models

The most successful approaches to the prediction of upcoming words are based on *n*-gram (or Markov) language models. These are very simple statistical models which are able to capture, to a remarkably large extent, the sequential word-by-word constraints of natural language. A typical *n*-gram model consists of the conditional probability space for occurrence of a word, given that *n* words have occurred immediately previously. Despite their lack of linguistic sophistication, trigram language models (where n=3) are used extensively in state-of-the-art speech recognition systems (*eg*. Jelinek, Mercer & Roukos, 1992). Adding linguistic information to the model (by first parsing the input, for instance) does not substantially increase accuracy. *N*-gram models (for English at least) seem to have the greatest success in predicting function words, which pose the greatest difficulty for automatic speech recognition due to the paucity of their acoustic signal.

6.3.2 The predictive probability

Recall from section 6.1.3 that the denominator of Bayes' rule, $P(c)$ (the probability of the data), permits predictions to be made about a new sample of data *before* it is observed. $P(c)$ is known as the *posterior predictive distribution*; it expresses the uncertainty about the new observation. The predictive distribution is written as:

$$P(c) = \int_{-\infty}^{\infty} P(c|\theta)P(\theta)d\theta$$

Note that the predictive distribution is dependent on both the likelihood and the prior probabilities. Because we treat a word's vector representation as a multinomial likelihood function, the predictive probability of a particular word vector $P(c_1 \ldots c_k)$ can easily be determined, by taking advantage of the fact that the predictive distribution of a Dirichlet posterior has a closed form. Stolcke (1994) shows that the above expression can be rewritten as:

$$P(c_1,\ldots,c_k) = \frac{n!}{c_1!\ldots c_k!} \cdot \frac{B(c_1+\alpha_1,\ldots,c_k+\alpha_k)}{B(\alpha_1,\ldots+\alpha_k)}$$

Given a set of prior weights α, the predictive probability for the vector representation of any lexeme in the vocabulary can be calculated accordingly.

It is now apparent that the ICD model (or at least the component for revising the semantic expectation) could also serve as a kind of *language model*, similar to those developed for applications in statistical NLP. Language models are essential components of accurate speech recognition software, spell-checking tools, statistical machine translation systems, and other language technology applications where word prediction is needed. Like *n*-gram language models, which estimate the probability of a word conditioned on *n* words of prior context, the predictive distribution as defined above enumerates the probabilities of word *vectors*, given a particular linguistic environment. Unlike *n*-gram models, which are limited to small contexts (typically n=3) for reasons of computational complexity and sparseness of training data, the size of the context that the Bayesian approach can use is unrestricted. Of course, the usefulness of this approach may well drop with the length of the previous context.

Although application to NLP may well prove useful, of more immediate interest to this thesis is whether the predictive distribution has any psychological reality. The obvious connection from predictive probabilities to psychological measures of word prediction is with *cloze probability* (*cf.* section 4.2). Is there a relationship between the cloze probabilities for the set of human-elicited completions of a particular sentence fragment and the predictive probabilities of their vector representations? Furthermore, is there any association between sentence constraint strength (as determined by the probability of the highest-cloze completion word) and the predictive probability of the completion word, across sentences? These questions are left open for future research.

6.4 Summary

The main contribution of this chapter has been to demonstrate how a parsimonious explanation of both context-independent and context-dependent lexical processing behaviour can be derived from environmental statistics. Building on the information-theoretic measure of contextual distinctiveness presented in Chapter 5, we developed a computational procedure for modelling the formation of semantic expectations from the linguistic context (the ICD model). The fact that language comprehension is incremental in nature indicates that semantic expectations should be available at any point before the end of a sentence; we tested the ability of the ICD model to account for the effort in processing words preceded by various amounts of context. Through simulation of a single-word priming experiment, we argued that ICD permits a novel, *information-oriented* interpretation of the lexical priming effect: priming is viewed as a *reduction* in the quantity of information conveyed by a target word about its meaning. We then proposed that contextual constraint effects could be described (and modelled) in exactly the same terms. This hypothesis was confirmed in Experiments 14 and 15 by reanalysing two eye movement studies from the literature. Constraint effects observed on eye movement variables were captured by the model. Finally, we assessed the generality of the ICD model by applying it to a corpus of eye movement data where contextual constraint was not explicitly manipulated. Although interpretation of the regression analyses was limited due to collinearity, when ICD – the theoretically preferable variable – was retained in preference to corpus frequency, it explained significant unique amounts of eye movement variance.

7.
Priming from Multi-word Contexts

The eye movement studies reported in Chapter 6 have shown that lexical processing effort is clearly modulated by the local linguistic context. But which aspects of context are responsible? Recent research on word recognition during sentence comprehension suggests that there are separate message-level and intralexical-level sources of contextual facilitation. This chapter critically examines the need for such a distinction, by drawing on evidence from both sentence priming and multiple-priming studies. We begin by discussing the results of sentence priming experiments that favour a message-level source of priming, and consider how the ICD model would account for the relevant data. In Experiment 17, the generality of the ICD model as a model of intralexical-level priming is tested through reanalysis of a multiple-priming study. The chapter concludes with a proposal for a new experiment which would address a strong prediction of the ICD model.

7.1 The source of sentence priming

In Chapter 4 we reviewed empirical research demonstrating the ability of the semantic context to modulate lexical processing behaviour. Recall that the constraining properties of the preceding linguistic context have been shown to influence the recognition of a sentence-final word, as well as to affect responses made to target words semantically related to the sentence-final item ('feature priming'). Both phenomena can be described as examples of *sentence priming*, which, generally speaking, refers to the process by which properties of the sentential context reduce the effort involved in processing subsequent words.

Clarifying the *source* of the sentence priming effect is an objective of current research (*eg*. Faust, 1998; Morris, 1994), and at least two distinct cognitive mechanisms have been proposed to underlie facilitation from the sentential context. The question posed is this: are context effects due to the existence of (a) semantic and/or associative relations between individual words in the sentence and the target word, or (b) a semantic relationship between the high-level, integrated representation of the meaning of the sentence context and the target? In other words, does sentence priming operate at the *intralexical* level, the *message* level, or some combination of the two? In this section we review the experimental evidence accumulated for both the intralexical-level and message-level hypotheses, and then examine whether the model of lexical processing effort (ICD) developed in Chapter 6 is suitable for addressing this question. We begin by briefly describing each position.

The message-level hypothesis maintains that it is the high-level representation of sentence meaning that is responsible for the majority of context effects in lexical processing. Word recognition is assumed to be speeded from the integration of conceptual, syntactic, and world knowledge into a representation of sentence meaning that is itself semantically related to the target word. Under this view, the semantic congruity effect described in section 4.2 (semantically compatible sentence-final words are easier to process than semantically incompatible words) reflects the compatibility of the target word with the message-level representation of the sentence context, and not to lexical relationships[1] between content words in the previous context and the target word. Semantically congruous completion words satisfy the constraints imposed by the message-level representation of the preceding context, whereas incongruous completions do not meet the constraints. What is crucial about the message-level standpoint is that priming is considered to stem from the integrated representation of the context – which is assumed to incorporate a variety of information sources – and not simply from the individual words in the context.

---

[1] By *lexical*, we mean semantic, associative, taxonomic, functional and other word-to-word relations that we assume to develop primarily through exposure to the distributional properties of words (or their referents) in the environment. There is evidence that the relation holding between *collocates* (words that co-occur more frequently than expected by chance) is also cognitively relevant, because collocations have been shown to support priming (McKoon & Ratcliff, 1992, Experiment 3). Note, however, that collocational, semantic and associative relations are often confounded (Moss, Ostrin, Tyler & Marslen-Wilson, 1995; Williams, 1996), making it convenient to use *lexical relation* as an umbrella term.

The opposing position to the message-level view is that sentence priming effects are intralexical in origin – they are due solely to pre-existing lexical relations holding between the words in the context and the target word (Forster, 1979; Tanenhaus & Lucas, 1987). Priming from multi-word sentence contexts is thus considered to be very similar to the standard account of single-word priming: the presence of lexical relationships between prime word(s) and target gives rise to facilitation, which is typically attributed to the automatic spread of activation between the memory representations of related words. As we shall see in section 7.1.2 below, adoption of the spreading activation metaphor for intralexical priming has compromised the conclusions that have been drawn about the role of lexical relations in sentence priming.

7.1.1 The empirical evidence

Of the experimental work attempting to ascertain the source of sentence priming, several researchers have capitalised on the assumption that a message-level representation of sentence meaning cannot be formed if the sentence context is presented 'scrambled'. The argument is that if scrambled (or *syntactically incoherent*) contexts like (1b) fail to induce priming, then any effects observed with the original (syntactically coherent) versions of the same sentences like (1a) could not simply be due to the existence of lexical relationships between individual word(s) in the sentence and the target.

(1) a. The author received much acclaim for his new *book*.
b. For author acclaim his much received new the *book*.
(examples from Simpson, Peterson, Casteel & Burgess, 1989)

A second line of investigation has manipulated the syntactic structure of the sentential context in order to alter its semantic relationship to the target word, while controlling the selection and order of the words in the context as much as possible. This syntactic manipulation is assumed to modify the message-level representation of the sentence. If priming effects are observed both from the original, unmodified sentence context and when the meaning conveyed by the entire sentence is different (resulting from the change in its syntactic structure), then the origin of sentence priming cannot be due solely to properties of the integrated message-level representation. However, if the modified context fails to induce priming, then this would constitute strong evidence for the message-level hypothesis.

7.1.1.1 Manipulating coherence

There is some evidence that the mere presence of lexical relationships between specific words in the sentential context and the target word is insufficient for priming to occur. Using the sentence priming paradigm and a rapid serial visual presentation (RSVP) procedure, Masson (1986, Experiment 1) found diminished target word facilitation when sentence contexts were scrambled (*ie*. sentence primes were syntactically incoherent), and subjects were required to make a lexical decision response to the target. There was no facilitation at all in the scrambled condition when the dependent variable was naming latency (Masson, 1986, Experiment 2), suggesting that it is the message-level representation of meaning that is largely responsible for the priming effect found with normal (syntactically coherent) sentence primes, and not the existence of lexical relationships between the target and individual words in the sentence context. Using similar experimental procedures, Simpson, Peterson, Casteel and Burgess (1989) and O'Seaghdha (1989) also manipulated the syntactic coherence of their sentence primes, and report a similar pattern of results.

Vu, Kellas and Paul (1998, Experiment 5) describe parallel findings using sentence primes containing ambiguous sentence-final words (homographs). Subjects named target words which were either related or unrelated to the particular meaning of the homograph that was biased by the preceding context; for example, the related and unrelated target words for (2) were **station** and **fly**, respectively.

(2) The soldier patrolled the *base.*

In this example, the context biases the MILITARY COMPLEX meaning of *base* over the BASEBALL usage, with predicted facilitation for **station** but not **fly**. Although contextual bias influenced response times when the sentence context was presented normally (*ie*. only **station** was primed), there was no bias effect when the words preceding the homograph were scrambled. Vu *et al.* suggest that the reason for the lack of context effect for the scrambled contexts was because "… it is not simply a matter of intervening words preventing priming but, instead, a matter of constructing a semantic representation to sustain meaning activation." (p. 992). Thus, their results support the message-level position. Vu *et al*. argue that context effects are normally due to message-level priming, but qualify this argument by conceding that "… lexical level priming is operative only when there is no coherent context and is restricted to the contiguous final word and target." (p. 992).

Finally, it appears that it is not syntactic coherence *per se* that is required for message-level priming; rather it is a more general notion of coherence that is the important factor. Simpson *et al.* (1989, Experiment 3) included a condition where the sentential context contained a strong associate of the target word and was syntactically well-formed but semantically anomalous. For example, unlike sentence fragment (3a), the semantically incoherent context (3b) failed to prime the target *cry* (compared with an unrelated context).

(3) a. The presence of the stranger made the baby *cry*.
b. The permit with the talent let the baby *cry*.

This finding suggests that for the scrambled context presentations discussed above, it is not simply the lack of syntactic coherence which prevents priming from occurring, but rather the inability of the human language processor to combine words in the order they are encountered into a meaningful representation. We normally encounter words in syntactically and semantically coherent units – phrases, sentences, and chunks of discourse – and it would be surprising if the language processor deals with incoherent sequences of words in the same way. It seems likely that semantic coherence in the sentence context is the minimum prerequisite for construction of a message-level representation.

In summary, the results of sentence priming studies that manipulate coherence – whether by scrambling the words in the sentence context or introducing semantic anomaly – offer support for the message-level hypothesis, by demonstrating that context effects cannot simply be due to the presence of lexical relationships between individual context words and the target.

7.1.1.2 Manipulating syntactic structure

Evidence against a *strict* message-level origin for sentence priming comes from studies where the lexical content of the context is kept virtually constant, while varying the syntactic structure. (Altering the syntax is assumed to substantially change the message-level representation of the sentence.) For example, using a RSVP naming task, Duffy, Henderson and Morris (1989) found equivalent priming effects for the sentence-final word in both (4a) and (4b):

(4) a. While talking to him the barber trimmed the *mustache.*
b. While talking to the barber she trimmed the *mustache.*
c. While talking to him the person trimmed the *mustache.*

Because the presence of the sentence priming effect seemed to be unaffected by changes in syntactic structure (*mustache* was primed both when *barber* was the agent of *trimmed* and when *barber* and *trimmed* were not in any syntactic relationship), and thus differences in the message-level representation of the previous context, Duffy *et al*. suggested that sentence priming is an intralexical phenomenon. But in (4c), where only one of the content words in the sentence context was related to the target (*ie*. *trimmed*), the context effect disappeared, leading Duffy *et al*. to argue that intralexical-level priming arises from the combination of at least two content words in the context that are related to the target.

However, results of follow-up research complicate these conclusions. In two eye-movement studies, Morris (1994) varied a number of sentence properties including the relatedness of the context's high-level conceptual representation to a target word. For instance, in Morris' Experiment 2 the message-level representation was manipulated by replacing function words in the stimuli, in order to drastically alter the functional relationships between referents in the sentence. For example, facilitation of the target noun *mustache* was predicted for (5b) compared with 'neutral' context (5a), but not (6b) compared with (6a):

(5) a. The friend talked as the person trimmed the *mustache* after lunch.
b. The gardener talked as the barber trimmed the *mustache* after lunch.

(6) a. The friend talked to the person and trimmed the *mustache* after lunch.
b. The gardener talked to the barber and trimmed the *mustache* after lunch.

Because the difference in the lexical content between the (a) and (b) sentences is equal for the two pairs, facilitation of *mustache* in (5) but not (6) would indicate that the message-level representation of the context is responsible for the effect. In terms of plausibility, it is more likely for a barber than a gardener to trim a mustache, and this pragmatic difference was reflected by the fixation time data: the difference in facilitation of *mustache* was significant only for (5). Morris concludes that these findings are inconsistent with an intralexical-level priming mechanism, since the target word needed to be semantically compatible with the message-level representation of the context in order for a sentence priming effect to be observed. However, the 'strong' message-level hypothesis was untenable, since priming was observed on the verb (*trimmed* in this example) from a related subject noun, *regardless* of syntactic structure, and hence independently of the high-level representation formed from the sentence context.

To recap, the empirical evidence considered thus far favours a combination of message-level and intralexical-level sources for sentence priming effects.[2] The hypothesis that contextual facilitation arises exclusively from intralexical sources is discounted by the results of Masson (1986) and Simpson *et al.* (1989), since the presence of lexical relationships between the target and individual words in the preceding context did not appear to be sufficient to reliably induce priming when the sentence context lacked syntactic or semantic coherence. The results of Morris' (1994, Experiment 2) eye movement study suggest a message-level source for facilitation of the target noun, but the priming effect observed on the verb makes a strict, message-level-only hypothesis untenable. Intralexical-level sources appear to have a limited role in sentence priming.

7.1.2 Predictions of the ICD model

We have seen that the evidence against an intralexical source of sentence priming has accrued mainly from studies where word order was disrupted. These studies have generally adopted the spreading activation metaphor in order to explain how priming occurs; they implicitly assume that activation spreads more or less *independently* from the cognitive representations of each word in the context, raising the activation levels of related words in semantic memory. Thus, spreading activation theory predicts priming from scrambled sentence contexts, which was not reliably attested. However, if activation flowing independently from the words in the sentential context is not the correct way to view intralexical priming, then conclusions drawn about the (limited) role of an intralexical-level source of contextual facilitation are premature. We propose that intralexical priming can be alternatively viewed in information-oriented terms; specifically, we suggest that the ICD model introduced in Chapter 6 is suitable for modelling contextual facilitation from intralexical sources. The ICD model makes three predictions about human processing behaviour in the sentence priming paradigm.

---

[2] Even this statement is an oversimplification, since we have ignored the important role for the lateralisation of brain function in sentence comprehension. The two cerebral hemispheres seem to represent sentence meaning in qualitatively different ways (Faust, 1998), which predicts hemispheric differences in sentence priming behaviour. Faust reports experimental evidence based on lateral presentation of stimuli that strongly suggests that the left hemisphere is specialised for message-level processing, whereas the right hemisphere is sensitive only to intralexical information.

First, ICD anticipates the sensitivity of lexical processing to syntactic coherence. Although manipulating the order of the words in the preceding context results in identical predictions of the size of the priming effect for a spreading activation-type model, the ICD model predicts different degrees of processing effort for target words preceded by normal and scrambled contexts – not because of any sensitivity to syntactic structure, but because it is sensitive to both temporal proximity and order. Recall that the ICD model was designed to model the incremental construction of semantic expectations. Because ICD's Bayesian update procedure is applied successively to the words in the sentential context, the contribution (or impact) of a word encountered early in the context diminishes as subsequent words are incorporated into the semantic expectation. This procedure implements the empirically-verified assumption that variability in processing effort is primarily dependent on the expectations derived from the 'closest' words to the target (Simpson *et al*., 1989, Experiment 3). The ICD model would therefore predict a priming effect from a syntactically coherent sentence context such as (7a), but not from an scrambled context like (7c), which is consistent with the empirical evidence: targets were facilitated (compared with a neutral context) in (7a) but not (7c) (Simpson *et al*., 1989).[3]

(7) a. The presence of the stranger made the baby *cry*.
b. The permit with the talent let the baby *cry*.
c. The made the of presence stranger the baby *cry*.

Second, the ICD model predicts the disruptive effect on priming from a semantically anomalous context observed by Simpson *et al*. (1989, Experiment 3). Even though sentence contexts such as (7b) were syntactically well-formed, *semantic* coherence was also required for priming to occur. A straightforward interpretation of this finding is that because the meanings of the words in the context bear little relation to each other, the processor cannot form useful semantic expectations, and is therefore unable to reduce uncertainty about the meaning of upcoming words. Thus, the ICD model is also predicted to be sensitive to semantic coherence, since the words in a semantically anomolous context will be ineffective cues for predicting the contextual behaviour, and hence the meaning, of the target word.

[3] In a *post-hoc* analysis, Simpson *et al*. (1989, Experiment 3) did find a priming effect in the scrambled condition when there was no gap between the strong associate and the target word (*eg*. … baby *cry*.). However, no comparable effect was found in their first two experiments, which used identical materials.

The third prediction relevant to sentence priming is that the ICD model will fail to distinguish between experimental situations where syntactic structure, but not lexical content, is manipulated. Because the Bayesian update rule is applied only to *content* words in the preceding context, ICD certainly would not be able to capture differences in priming behaviour of the type described by Morris (1994, Experiment 2). Recall that the priming effect disappeared when the function words in the context were replaced in order to alter the syntactic relationships between critical lexical items – the order of the content words remained the same.

In summary, conclusions about the relative importance of intralexical-level and message-level sources of contextual facilitation in sentence priming are compromised when an alternative model of intralexical priming – one that does not rely on the spreading activation metaphor – is adopted. Because the ICD model is predicted to account for the processing differences observed between coherent and either syntactically or semantically incoherent sentence contexts – without purporting to construct or access a high-level conceptual representation of the sentence – it offers a superior explanation of sentence priming than spreading activation theory. An intralexical source for priming is not as easily dismissed when ICD replaces the traditional spreading activation model. ICD fails, however, in predicting attested differences in lexical processing effort resulting from manipulating syntactic structure. No matter how much contextual facilitation can be attributed to intralexical sources, a role for priming from the message-level representation of the sentential context must still remain (at least for normal sentence comprehension). It appears that the ICD model is best considered as capturing the variability in processing effort attributable to intralexical-level facilitation.

Recall that in Chapter 6 we presented the ICD model as a way of objectively quantifying the constraining effect of the lexical items in the local linguistic context using the tools of information theory. It should now be clear that ICD provides exactly the same explanation for sentence priming as it does for single-word priming and contextual constraint effects (a reduction in the amount of information conveyed by the target word about its meaning), where the influence of the local context is viewed as the incremental revision of the processor's expectations about the meaning of upcoming words. Although the effects of contextual constraint and priming are conventionally assumed to be 'semantic', the only information available to the ICD model about the meaning of a word is what is latent in its distributional pattern of use.

7.2 The multiple-prime advantage

The conclusion that intralexical-level processes are insufficient for explaining priming from multi-word contexts may be premature. Construction of a message-level representation of the linguistic context would not appear to be necessary to support priming if one compares sentence priming studies with experiments looking at the effects of *multiple* prime words on target word recognition. In the latter paradigm (also known as *summation priming*), all of the prime words are related to the target, and although they are typically presented sequentially, they do not form a syntactically coherent, natural language-like context for the target word.

Past research in summation priming has typically created multiple-prime contexts for the related condition using members of the target word's semantic category; such primes would be expected to facilitate recognition of the target if used individually in a conventional single-word priming paradigm. Results of multiple-priming studies support an intralexical source for context effects, because several researchers (*eg*. Brodeur & Lupker, 1994; Balota & Paul, 1996) have demonstrated a *multiple-prime advantage*: two or more related prime words triggered a larger overall priming effect than a single related prime. Priming is *additive*, and is typically construed as the increased facilitation resulting from activation converging on the target word from multiple sources. In this section, we investigate the ability of the ICD model to explain the multiple-prime advantage, and thus provide an alternative to the spreading activation account. This phenomenon constitutes a further test of the generality of the ICD model, which would strengthen the information-oriented view of lexical processing effort advocated in this thesis. If one accepts that the function of the human sentence processor is to integrate the meanings of the words in an utterance into a high-level conceptual representation, then multiple-priming behaviour is of interest because it is not obvious that such a message-level representation could be formed from an unstructured sequence of prime words. But if similar processes underlie *both* sentence priming and multiple-priming, then the ICD model should succeed in capturing both types of contextual facilitation.

7.2.1 Previous research

Brodeur and Lupker (1994) employed a multiple-priming paradigm that compared the effects of four related primes with a single related prime, using stimuli consisting

of category exemplars (see Table 7-1 for sample materials). Results from lexical decision, but not naming tasks, indicated a reliable multiple-prime advantage. An interesting finding was that lexical decision responses in their four-prime unrelated condition were *slower* than responses in the corresponding one-prime condition, suggesting that the multiple-prime 'boost' was actually due to inhibition in the unrelated condition, not facilitation. Brodeur and Lupker conclude that the conceptualisation of priming as involving a flow of activation between memory representations is too simplistic to explain their results. They also suggest that it is higher-level 'post-lexical' processes – which they consider lexical decision, but not naming to be sensitive to – that are influenced by the number of related primes preceding the target word. Thus, Brodeur and Lupker's work illustrates a pattern of contextual influences on lexical processing effort that cannot be adequately explained using the spreading activation metaphor.

Balota and Paul (1996) were also interested in how multiple primes (again, construed as independent sources of spreading activation) could influence target word processing. Using two-word contexts, they separately manipulated the relatedness of each prime to the target word; this procedure allowed additive priming effects to be closely measured. In order to address objections that any attested effects were actually due to 'post-lexical' strategies, they employed three different experimental tasks: lexical decision, naming and relatedness judgements.

Balota and Paul note that previous multiple-priming studies have confounded prime-target relatedness with relatedness *between* the prime words. For example, in the presentation <*copper*, *bronze*, **metal**> the two prime words are semantically related to each other, as well as to the target word **metal**. This leaves open the question of whether (employing the spreading activation metaphor) each prime

Table 7-1. Sample Materials from Brodeur and Lupker (1994).

Condition	Prime-1	Prime-2	Prime-3	Prime-4	Target
Four-prime					
Related	snow	thunder	hail	sunshine	rain
Unrelated	bee	flea	roach	spider	sofa
One-prime					
Related	–	–	–	robin	crow
Unrelated	–	–	–	fraud	knee

Note: Actual combinations of prime and target words were not provided by Brodeur and Lupker; these example items were constructed according to the criteria detailed in their Methods.

independently activates the semantic category target, or the first prime enhances the activation level of the second prime, which in turn activates the target to a greater degree. Balota and Paul (1996, Experiment 1) address this concern by manipulating *inter-prime* relatedness; their design included a condition where both primes were related to a homographic target word, but not obviously related to each other (*eg.* <*kidney*, *piano*, **organ**>). For these stimuli, the first prime should not affect the activation level of the second, and hence the presence of additive priming could not be due to inter-prime relationships. The results of Balota and Paul's Experiment 1 indicated equal-sized additive priming effects for both types of target word (category label and homograph), leading the authors to argue that previous reports of the multiple-prime advantage (*eg.* Brodeur & Lupker, 1994) cannot be the result of inter-prime facilitation.

However, using a relatedness judgement task (which required subjects to indicate if either or both of the prime words were related to the target), Balota and Paul (1996, Experiment 6) report a different pattern of results for their ambiguous and unambiguous target stimuli: there was no additive priming effect for the homograph targets. For these items, there was minimal influence on response times from encountering a second prime word that was related to a different aspect of target word meaning than the first prime. Based on this pattern of results, Balota and Paul argue that the relatedness judgement task forced subjects to access meaning, not merely lexical-level (*ie.* non-semantic) relations between prime and target words.

Balota and Paul conclude that the context effects found in Experiments 1-5 were not due to processing at a semantic or conceptual level, but were the result of intralexical-level facilitation. Because the pattern of priming did not differ for target words that mapped to multiple concepts (*eg.* **organ** $\rightarrow$ BODY PART $\vee$ MUSICAL INSTRUMENT), compared to unambiguous targets (*eg.* **metal**), there did not appear to be any semantic-level involvement. In other words, the experimental evidence was not consistent with models that postulate distinct conceptual representations corresponding to a single lexical representation. Their results are compatible with the ICD model, however, which implicitly recognises a single, 'lexico-semantic' level of representation. Earlier in this thesis we argued that co-occurrence vector representations encode semantic as well as associative / collocational information (*cf.* section 3.3.3), precluding the need for distinct mechanisms or levels of representation in explanations of semantic priming. Balota and Paul's results lend further support to our claim. If the source of contextual facilitation in multiple-

priming is at the intralexical level, then the ICD model should be able to account for both additivity and the insensitivity to inter-prime relatedness demonstrated by Balota and Paul. In Experiment 17 we submitted Balota and Paul's (1996, Experiment 1) materials to a computational reanalysis, in order to assess the ICD model's ability to capture the pattern of contextual facilitation revealed using the multiple-priming paradigm.

7.2.2 Experiment 17

The main goal of Experiment 17 was to test the generality of the information-oriented ICD model as a model of facilitation from multi-word contexts. Unlike sentence fragments, multiple-prime contexts are completely unstructured, and it is difficult to see how a high-level conceptual representation could be constructed. Consequently, if ICD captures the fundamental processes that underlie the variation in lexical processing effort observed with sentence primes, and if these processes also drive the priming effects reported for unstructured multi-word contexts, then the ICD model should additionally account for multiple-priming effects. Successful simulation of the pattern of human behaviour reported by Balota and Paul would provide additional support for ICD as a model of intralexical-level facilitation.

The computational reanalysis had two further goals; namely to test Balota and Paul's claims regarding (a) the *additivity* of multiple primes, and (b) the influence of the *temporal proximity* of the related prime to the target word.

In their first experiment, Balota and Paul demonstrated that the multiple-prime advantage was additive: the facilitation obtained in the two-related-primes condition (RR) was equivalent to the sum of the facilitation for the one-related-prime conditions (UR and RU). (See Table 7-2 below for sample stimuli). Because they found evidence for simple additivity using a range of prime presentation durations, and both lexical decision and naming as response tasks (Balota & Paul, 1996, Experiments 1-5), the authors state that "… we believe that contextual constraints can produce simple additive influences on target processing." (p. 839). In terms of the ICD model, two related prime words would need to constrain the processor's expectations about the meaning of the target to a greater degree than a single related prime in order to simulate the multiple-prime 'boost'. Although this seems intuitive for the semantic category stimuli (*eg.* <*copper*, *bronze*>), it is not at all obvious that the prime words for the homographic targets (which are not related to each other: *eg.* <*kidney*, *piano*>) will induce the same effect.

Also of interest in the human experiments was the difference in the size of the priming effects between the UR and RU conditions. The temporal proximity of the related prime word to the target seemed to be important, since larger effects were obtained in the UR condition, over most of their experiments.[4] Balota and Paul suggest several possible interpretations of this finding, including "disruption of priming by an intervening word" (p. 832). It would be desirable if the ICD model also captured this proximity effect, since an intervening unrelated word should intuitively have an adverse effect on the processor's expectations about the meaning of the upcoming target word. The proximity of the related prime is predicted to modulate the size of the simulated context effect.

7.2.2.1 Method

The design was identical to that of Balota and Paul's Experiment 1. This was a 2 × 4 mixed factors design, with Type of Target (Homograph, Category Label) as the between-items factor, and Prime Type (RR, UR, RU, UU) as the within-items factor.

Preparation of the lexical stimuli was very similar to the procedure carried out for the computational reanalyses reported in previous chapters. Stimuli were first converted to their canonical forms, and items containing target or related prime words that did not meet the 25-occurrence frequency threshold were removed. As in earlier simulations, this was done to ensure an adequate level of reliability for the co-occurrence statistics. Unrelated prime words that failed to meet the frequency threshold were replaced with unrelated primes chosen randomly from the set of discarded items. From the 106 original homograph items, 69 could be used in the simulation. Out of the 94 original category stimuli, only 39 met the frequency criterion. (See Table 7-2 for sample materials).

Next, vector representations for the critical stimuli were created from the BNC-spoken using the 'standard' parameter settings: the window size was ±5 words, and the 500 most frequent content words in the BNC-spoken served as context words. Finally, we calculated ICD values for each target word when preceded by each of the four Prime Type conditions, using the same parameter settings employed in Chapter 6's simulations (total prior weight=2,000, sample size=500).

[4] Note that the difference in mean RT between the RU and UR conditions rarely reached statistical significance in a single experiment; however, after collapsing over the RT data from Experiments 1-5 this difference was highly reliable.

Table 7-2. Results of the Simulation of Balota and Paul (1996, Experiment 1), with Mean Lexical Decision Response Times (RT) and Amount of Priming (Priming).

	Example Stimuli			ICD	RT	Priming
Condition	Prime-1	Prime-2	Target	(bits)	(msec)	(msec)
Homograph targets						
RR	game	drama	play	0.895	601	34
UR	lip	drama	play	0.970	618	17
RU	game	tuna	play	0.932	630	5
UU	lip	tuna	play	1.011	635	
Category label targets						
RR	hate	rage	emotion	1.095	606	34
UR	author	rage	emotion	1.151	616	24
RU	hate	design	emotion	1.114	627	13
UU	author	design	emotion	1.193	640	

Note: R=related prime, U=unrelated prime.

7.2.2.2 Results and Discussion

Analogous to the simulations described in Chapter 6, facilitation was interpreted as a *reduction* in the amount of information conveyed by a target word about its contextual behaviour in one of the Related prime conditions (RR, RU and UR), compared with the UU (two-unrelated-primes) condition. Facilitation was apparent for all three Related conditions. The size of the context effect was 0.110 bits for the RR condition, 0.041 bits for the UR condition, and 0.079 bits for the RU condition. The differences in mean ICD value were verified by an analysis of variance, which revealed a main effect of Prime Type, $F(3,306)=40.53$, $p<0.001$. There was no reliable effect of Target Type, $F(1,102)=2.26$, $p=0.135$, and no evidence for a Prime Type × Target Type interaction, $F(3,306)<1$. Table 7-2 displays the relevant mean ICD values for the separated homograph and category label stimuli.

The pattern of results was closely comparable to the human data. As expected, the strongest context effect was observed in the RR condition, which was larger than the effects in both the UR and RU conditions. This finding replicates the multiple-prime advantage reported by Balota and Paul. Additivity was also present in the simulation results. The sum of the facilitation obtained in the UR and RU conditions (0.079 + 0.041 = 0.120 bits) was nearly identical to the effect size obtained when both primes were related (0.110 bits for the RR condition). The reduction in information conveyed by a target word preceded by two related primes was a

simple additive function of the reduction in information resulting from separate presentations of each prime.

In contrast to the human data, the context effect for the RU targets was *larger* than for the UR targets. At first glance, it appears that (at least for the current ICD model parameter settings) the temporal proximity of related prime to target did not produce the anticipated effect in the simulation, but rather the opposite.[5] This difference between the RU and UR conditions was statistically reliable: planned comparisons (with suitable alpha corrections) confirmed that all four conditions differed significantly from one other, at the α=0.05 level of significance.

The presence of the multiple-prime advantage for both unambiguous *and* ambiguous target words (*ie*. the lack of Prime Type × Target Type interaction) indicates that prime words did not have to be related to each other in order for the model to simulate the multiple-prime 'boost'. This is a noteworthy finding: although the co-occurrence patterns for pairs of prime words such as <*kidney*, *piano*> would clearly be expected to differ to a much greater extent than pairs such as <*copper*, *bronze*> (corresponding to differences in semantic relatedness), the ICD model was still sensitive to the fact that individually, the vector representations for *kidney* and *piano* are useful sources of information for predicting the contextual behaviour of the target word **organ**.

In summary, the ICD model was largely successful in capturing the pattern of contextual facilitation obtained by Balota and Paul in their multiple-priming study. Although modulation of priming by the temporal proximity of the related prime to the target word was not attested, the multiple-prime advantage – a larger context effect for two related primes compared to one – and simple additivity were present in the simulation.

[5] The reason for the larger simulated priming effect for the RU condition is likely due to the predicted size of the *independent* (single-word) priming effect being larger for the set of Prime-1 words than the set of Prime-2 words. Using the UU condition as the unrelated baseline, the effect size was 0.125 bits using only the Prime-1 set and 0.070 bits using the Prime-2 set. It appears that for the ICD model, Balota and Paul's set of Prime-1 words is substantially 'better' than the second set. The distributional characteristics of the Prime-1 words diverge enough from those of the Prime-2 words to give rise to this difference in simulated priming.

7.3 A unified view of sentence priming and multiple-priming

Taken together with ICD's predictions for sentence priming behaviour, the ability of the information-oriented model to explain multiple-priming effects provides strong support for Chapter 6's claim that the ICD model captures the semantic constraint imposed by multi-word contexts. Multiple-priming effects cannot easily be attributed to message-level processes, because is not clear that a meaningful conceptual representation can be formed from a set of prime words (especially when the primes are not related to each other). The fact that ICD offers a single explanation for both multiple-priming and sentence priming behaviour suggests that similar processes are responsible for both types of contextual facilitation.

This conclusion – common processes underlying both types of priming – leads to an interesting prediction. If priming from a sentence fragment is primarily due to the presence of lexical relationships between the target and content words in the fragment, then priming contexts constructed using *only* those content words should also produce facilitation. In other words, the ICD model makes the strong prediction that human lexical processing behaviour will be similar for primes presented *either* as syntactically coherent sentence fragments *or* as simple sequences of content words.

This hypothesis could be tested with a variant of the sentence priming paradigm used in previous research (*eg*. Masson, 1986), but involving two tasks, both requiring subjects to pronounce the final word: (a) sentence fragments are presented sequentially using RSVP, and (b) only the content words from those fragments are displayed. This would constitute the Related condition. Unrelated primes would be formed from randomly reassigning contexts to target words. Priming would be measured as the difference in naming latencies between the Unrelated and Related conditions for the same target word. The strongest prediction is that the type of context (sentence fragment or sequence of content words) will fail to interact with Relatedness; although perhaps this prediction is too strong. Even if the priming effect is larger for the sentence fragments, it should still be present for the 'multiple-prime' contexts. The presence of a Relatedness effect, but absence of an interaction, would provide clear support for the hypothesis that both sentence priming and multiple-priming operate at the intralexical level. More importantly, such a result would further validate ICD as a context-dependent model of lexical processing effort.

Although we are arguing for a unified view of sentence- and multiple-priming, an open question remains concerning the human behavioural data. Why do multiple-priming contexts of the type tested by Balota and Paul (1996) support facilitation, but scrambled sentence contexts do not? On the surface they seem quite similar, since both types of context lack syntactic coherence (and the contexts for Balota and Paul's homograph targets could also be described as lacking semantic coherence). How can these findings be reconciled? One possibility is that the two tasks are different enough that subjects employ qualitatively different strategies. However, this is not a satisfying explanation. A more likely answer is that the prime words used in multiple-priming studies are simply better predictors of the meaning of their designated target words than the content words found in sentence priming contexts. In terms of the ICD model, the former type of context simply allows more precise semantic expectations to be constructed about the meaning of the target word, which reduces the effort involved in processing that word.

Human lexical processing is sensitive to contextual factors such as syntactic and semantic coherence, the presence of intralexical relationships between target and context words, and the temporal proximity of the words in these relationships. We have demonstrated that a substantial portion of sentence- and multiple-priming phenomena can be captured by a model of processing effort derived from the distributional information inherent in the linguistic environment.

7.4 Summary

The goal of this chapter was to provide further support for the ability of the ICD model to explain context-dependent variability in lexical processing effort, by applying the model to the sentence priming and multiple-priming paradigms. After reviewing contemporary research in sentence priming, we considered how the ICD model, given its simplicity, could account for the relevant behavioural data. By using the ICD model to address the issue of the origin of sentence priming, we were led to conclude that an intralexical-level source of facilitation has been prematurely dismissed by previous research that adopted the mechanistic view of priming as a spread of activation. Further evidence that the ICD model captures the semantic constraint imposed by multi-word contexts was provided by Experiment 17's reanalysis of Balota and Paul's (1996) multiple priming study: human performance was effectively simulated by the ICD model. Chapter 7 concludes by arguing that

the ICD model offers a unified explanation of both multiple-priming and sentence priming behaviour, which suggests that common cognitive processes underlie both types of contextual facilitation.

8.
Conclusions

The primary aim of this thesis has been to contribute towards an understanding of *lexical processing effort*. In order to address this goal, we have followed a virtually unexplored path, which involved forging a link between measurable properties of a word's linguistic environment and processing difficulty. We have argued that the simple distributional information contained in a large corpus of language output provides a rich source of information about word meaning, and more importantly, allows the formalisation of a computational account of language comprehension as a process of building expectations about meaning.

Paralleling recent work on understanding cognition using the methods of rational analysis, we have attempted to provide a *purposive* (or, to use Marr's [1982] term, *computational-level*) explanation of lexical processing effort. That is, we have tried to explain variability in processing effort in terms of the ultimate *function* of the lexical processor, as opposed to clarifying the cognitive architectures and algorithms used by the processor (the *mechanistic* style of explanation). We proposed that a computational-level explanation of lexical processing effort involves satisfying three subgoals: elucidating the function of the language processor, developing a computational model of the cost incurred in implementing the function, and successful modelling of the relevant experimental data.

Achieving the first subgoal was straightforward. We believe that the fundamental goal of human language comprehension is the recovery of meaning from an utterance, which entails recovering the meanings of the individual words in the utterance. In order for this process to be *efficient*, we suggested that a useful strategy would be for the processor to form expectations about the meaning of upcoming words.

The proposed expectation-building strategy is consistent with one premise of the rational analysis approach to cognition, namely that human cognitive behaviour is optimally adapted to its environment. In addition to increasing processing efficiency, forming expectations about the meaning of upcoming words would be beneficial for dealing with noisy input and ambiguity. Both phenomena are prevalent in human linguistic communication, and it is reasonable to assume that the processor has developed strategies to maximise the chance of communicative success in the face of obstacles such as these.

The expectation-building assumption led to the hypothesis that a non-trivial portion of the variability in behavioural measurements of lexical processing effort could be explained as the ability of the meaning of upcoming words to meet prior expectations. The 'fit' between expected meaning and actual meaning we suggested would be reflected by processing cost – the better the match, the less processing cost incurred when recovering the meaning of the forthcoming word. This hypothesis can also be phrased in terms of predictability – the more predictable the meaning of an upcoming word, the less effort required to process a word that conveys that meaning.

We then developed a computational model of this processing cost by applying the tools of information theory to lexical representations derived from the distributional information inherent in the linguistic environment. The fit between expected meaning and actual meaning was estimated as the amount of *information* conveyed by a word about its contexts of use. We showed that the expression for this quantity permits the contribution of prior knowledge to be mathematically incorporated.

Finally, the third subgoal was satisfied by a number of empirical studies. These studies largely took the form of computational simulations and reanalyses of published research concerned (perhaps implicitly) with 'semantic' influences on processing effort. In all cases we believe the behavioural data were adequately explained by our model of processing effort, and in several instances the modelling results invited serious reinterpretation of the original conclusions.

Rather than summarising the results of the various computational experiments conducted here on a chapter-by-chapter basis, we will first recap the main contributions of the thesis, and then discuss the role of distributional information in explaining the relevant behavioural phenomena.

8.1 Contributions

We have employed distributional information as the basis of two theoretically distinct, though mathematically related approaches to modelling psycholinguistic phenomena. In the first part of the thesis, we constructed a semantic space model from co-occurrence statistics; we considered this to be a type of *representational* model, because words are assigned distinct vector representations, and the explanatory power of the model rests on the hypothesis that representational similarity is relevant to lexical processing behaviour. This is more or less the conventional way to view the role of distributional information, and has proved to be a fruitful one. We extended this approach to account for context-dependent phenomena, by weighting the similarity measure according to statistical characteristics of the vector representations for the words in the immediate context. Although this model (coined the Contextual Relevance Model) proved to have some nice properties, it also has some notable limitations. The main drawback of the representational approach is that predictions of processing effort cannot be made for a *single* word; obviously, measures of representational similarity require *two* or more entities.[1]

The second half of the thesis supplants the representational approach to understanding lexical processing effort by a model that more directly instantiates the strategy of building expectations about meaning. Although the same source of distributional information is exploited by both approaches, we feel the latter permits a more parsimonious account of the behavioural data for two principal reasons. First, behavioural predictions are available for the processing of words encountered either in isolation or in a connected linguistic context (see next section). Second, consistent with what is known about on-line language comprehension, the ICD model (crudely) implements the construction of semantic expectations as an *incremental* process.

[1] At least one measure has been proposed that can be computed for a single word: this is the number of other words falling within a certain radius of a word's position in high-dimensional semantic space (Burgess & Lund, *in press*). Although 'semantic neighbourhood density' is certainly quantifiable for any word in the lexicon, it is decidedly lacking in explanatory power – *why* this measure should be relevant to normal lexical processing behaviour.

8.1.1 An integrated model of context-independent and context-dependent behaviour

One of most important contributions of this thesis concerns the ability of the [I]CD model to provide an integrated account of the difficulty of processing a word in the absence *or* the presence of context. Even though one normally encounters words in a connected linguistic context, the majority of work on lexical processing conducted in the psychological laboratory has investigated the recognition of isolated words. This paradigm is attractive due to the relatively tight control the experimenter has on variables extraneous to the factor of interest, but it is not completely clear whether explanations of context-independent processing behaviour generalise easily to context-dependent behaviour.

We have shown that with *no* architectural modifications, the same information-oriented model can explain variability in lexical processing effort observed in or out of context. This was accomplished solely through specification of the source of prior knowledge about meaning that is an integral part of the model. For the context-independent case, we assume an uninformative source of prior knowledge; for the context-dependent case, the prior incorporates distributional information from the words in the preceding context. Consistent with our purposive explanation of processing effort, we consider the strategy of forming expectations about meaning to apply to both situations; for isolated words, these expectations are relatively uninformative – they are in some sense 'neutral.' Under the [I]CD model, context-independent and context-dependent behaviour are indistinguishable in *qualitative* terms; they are predicted to differ (and do differ) *quantitatively*, as reflected by measures of processing difficulty.

Although a number of corpus-derived statistical measures (such as pointwise mutual information, conditional probability and simple co-occurrence frequency) are sensitive to the contexts in which a word is used, and such measures have been investigated for their ability to predict human behaviour (*eg*. Lapata, McDonald & Keller, 1999; McKoon & Ratcliff, 1992), [I]CD differs from these other corpus-derived measures in one important respect. [I]CD makes predictions about the processing of a word presented in isolation or in context, whereas measures such as mutual information are only relevant for pairs or sequences of words. This novel feature clearly distinguishes [I]CD from other corpus-based statistical models.

The preceding discussion brings us to an interesting conjecture. Perhaps there is no circumstance where 'true' context-independence holds. As in the single-word

priming paradigm, stimuli are presented sequentially in the typical isolated word recognition task. Ostensibly, there is no relationship between items, but preceding stimuli could provide a minimal, though semantically incoherent context for each target word. The ability of the lexical processor to make use of a semantically coherent preceding context is what drives the semantic priming effect, so in effect the two laboratory situations may not differ as much as one thinks.

8.1.2 A unified view of semantic context effects

We have proposed that the strategy of using previous context to form expectations about the meaning of upcoming words is a general (and in some sense, automatic) process. If the processor does adopt such a strategy, its effects should be manifest under any number of laboratory paradigms where semantic context is manipulated. The prediction is that the ICD model should explain some portion of the variability in lexical processing effort observed across different experimental paradigms and tasks. This is precisely what has been found. Single-word lexical priming, multiple-priming, sentence priming and contextual constraint effects are all accommodated by the ICD model, and are all explained as a *reduction* in the quantity of information conveyed by the target word about its contextual behaviour. The results of studies examining the influences of semantic context on processing effort were captured using a single, general model. Because human performance was established using a variety of measurement techniques in the original experiments – lexical decision response times, naming latencies, and eye movement data – the appeal of the information-oriented approach is further strengthened.

In Chapter 6, we suggested that a variety of semantic context effects could be subsumed under the notion of contextual constraint. The standard definition of contextual constraint is phrased in terms of predictability: the greater the degree of constraint, the more predictable the upcoming word. This is closely related to our description of the strategy of building expectations about meaning.

Finally, we have demonstrated that the ICD model offers an objective alternative to the traditional spreading activation model of contextual facilitation. ICD explains priming in a very different way – not in the mechanistic terms of a forward spread of activation within semantic memory – but as a reduction in the amount of information conveyed by a word about its contexts of use.

8.1.3 Contextual distinctiveness

Another major contribution concerns the development of a novel measure of environmental complexity, *contextual distinctiveness* (CD). Although CD is simply the uninformative-prior counterpart of the ICD model, it also functions as a descriptive variable, providing a new way to quantify the distributional differences between words. Formulated in information-theoretic terms, CD states that words that occur in distinctive (or restricted, or constrained) contexts tend to convey more information about their contexts of use than words that occur in relatively unrestricted (or unconstrained) contexts.

CD is an *objective* measure – it is easily calculated from a record of language output – that we have shown to correlate with several other *subjective* 'semantic' variables. It was found to predict one lexical property remarkably well: degree of semantic ambiguity (roughly, the number or range of meanings conveyed by a word) is significantly correlated with a word's CD value. The more specific the meaning evoked by a word, the more distinctive its contextual behaviour. This reflects the close correspondence between context and meaning advocated by Cruse (1986); certain meanings are associated with certain linguistic contexts of use, with the consequence that words mapping to more than one meaning tend to appear in a greater variety of contexts than words mapping to a single meaning. This finding has clear implications for word recognition studies where ambiguity is manipulated as an experimental factor. It is apparent that for many of these studies, CD will be a confounding variable, and patterns of results may well change once CD is controlled.

8.1.4 Rethinking the word frequency effect

CD is also naturally correlated with corpus frequency: frequent words tend to occur in a broad range of contexts, whereas the contextual behaviour of rare words tends to be much more constrained. There are good reasons for computing both CD and frequency of occurrence, as both are simple statistics available from the linguistic environment that have the potential to influence lexical processing behaviour. We have claimed that CD, not corpus frequency, is the more psychologically relevant predictor of processing effort, with the consequence that the large body of research supporting the word frequency effect could be reconstrued with CD as the explanatory variable. CD is at least as convincing a predictor as frequency, but

holds much more theoretical interest. By explicitly attaching importance to the properties of a word's contexts of use – not just the event of word occurrence itself – CD emphasises the natural interdependence between words in real language.

The most compelling evidence for the superiority of the information-theoretic [I]CD model is provided by Experiment 15's reanalysis of Altarriba *et al.* (1996). The reanalysis showed that a single variable could capture independent effects of word frequency and contextual constraint on lexical processing behaviour. The hypothesis that CD subsumes frequency as a predictor of lexical processing effort certainly needs to be investigated further.

8.2 Coverage

We have presented empirical evidence showing that a wide range of human behavioural data can be captured using the distributional information inherent in the linguistic environment. The conclusions of each study are summarised below:

Isolated word recognition

Two experiments demonstrated CD to be a psychologically valid predictor of processing effort in the visual lexical decision task. In the first experiment, CD explained a significant amount of response time variance; however, corpus frequency was a potential confounding variable. A comparable effect was found in the second experiment using pairs of words closely matched for frequency, although this effect was revealed as an interaction between CD and speed-split participant group. The question left outstanding – one that has serious repercussions for the large body of word recognition research that supports frequency as an explanatory variable – is whether the apparent effect of frequency can be better explained in terms of between-word differences in the amount of information a word conveys about its contexts of use.

Naming performance by Alzheimer's patients

Reanalysis of Chenery *et al*'s (1996) study of the picture naming errors made by Alzheimer's patients showed that CD was a significant predictor of naming performance on a by-item basis. High-CD picture names caused more difficulty than picture names with a low CD value. This suggests an explanation of the semantic impairment typical of Alzheimer's dementia that is consistent with our account of

word recognition performance by normal subjects: the degree of naming difficulty is greater for objects whose names convey the most information about their linguistic contexts of use.

Semantic similarity judgements

Our measure of the similarity of vector representations, Contextual Similarity, significantly correlated with a published set of semantic similarity ratings (Miller & Charles, 1991), and with a newly elicited set of ratings. These correlations supported our working hypothesis that useful information about the meaning of a word is latent in its distributional pattern of use.

Single-word lexical priming

The results of Moss *et al.*'s (1995) semantic priming experiment were successfully simulated using distributional information as a type of representational model. We showed that both categorial and functional relations are implicit in co-occurrence statistics. In this case, distributional similarity was assumed to correspond to variability in processing effort. By replicating the 'associative boost' in semantic priming observed by Moss *et al.*, the simulation questioned the need for distinct levels of representation or separate priming mechanisms in explanations of semantic and associative priming.

Feature priming

Computational reanalysis of Moss and Marslen-Wilson's (1993) feature priming study was not initially successful. Although intuitively appealing, the assumption that priming of a property target could be simulated by the relevance-weighted representational similarity between prime and target words was not supported by the human data. However, we advanced an alternative explanation for Moss and Marslen-Wilson's counterintuitive findings, and tested this prediction using the representational model. By modelling priming as the influence of the words in the context preceding the (designated) sentence-final prime, the pattern of human results was approximated. The simulation results demonstrated that the feature priming effect may actually be a function of the 'global' context, not just the designated sentence-final prime word. The simulation also had methodological implications: Moss and Marslen-Wilson's pre-testing of their sentence context materials (using subjects' intuitions) appeared to be inadequate. More objective tests for the presence of lexical relationships between the words in a sentence

context and a target word are supplied by measures derived from distributional information.

Contextual constraint

The influence of contextual constraint on eye movements during reading was effectively captured by the ICD model. Reanalyses of two eye movement experiments (Altarriba *et al*., 1996; Schustack *et al*., 1987) indicated that the effect of a semantically constraining context on lexical processing effort – as revealed by various measures of eye movement behaviour – can be modelled as a reduction in the amount of information that a word conveys about its context of use. Applying the model to a corpus of eye movement recordings (arguably a stronger test of its ability to explain normal reading processes) corroborated these findings. Although multiple regression techniques were unable to reveal ICD as a significant predictor independent from corpus frequency, when frequency was not included in the regression equation, ICD accounted for significant amounts of eye movement variance.

Multiple-priming

The ICD model closely simulated the results of Balota and Paul's (1996) multiple-priming experiment showing that presentation of two related prime words produced an additive priming effect. This reanalysis was particularly interesting because the model also captured the additive effect attested using pairs of primes that were unrelated to each other.

8.3 Semantics

A question which has arisen repeatedly during the course of this project, though not yet committed to discussion, concerns the status of *semantic representation* in psychology. Is the concept of 'semantic-ness' simply an epiphenomenon of objective, observable properties of the linguistic environment? Although we have used 'semantic' to describe the nature of the information latent in distributional statistics, and have employed terminology such as 'semantic influences', 'semantic context' and 'semantic priming', it has become clear that the term 'semantic' has failed to exhibit any explanatory power. In every instance that we have used the term, it has merely been as a convenient shorthand for a phenomenon that we have subsequently

attempted to explain in terms of distributional information. Moreover, several of the simulations have shown that 'semantic' often has too narrow a scope, and we have struggled to find a suitable replacement, employing 'lexico-semantic' or simply 'lexical'. It is apparent that much of what has been described as 'semantic' in psycholinguistic research can just as easily be described by recourse to the statistical information present in the linguistic environment.

The results of Experiment 2B (Chapter 3) suggest that semantic space models also contain information of a syntactic nature. Although co-occurrence data involving function words were excluded, the lower correlation obtained between cross-category similarity judgements and Contextual Similarity indicates that syntactic constraints are also implicit in co-occurrence vector representations. It appears that the most appropriate label to apply to the content of distributional statistics is 'contextual'. Whatever the label used, this thesis has shown the importance of distributional information for modelling language behaviour.

8.4 Limitations

It must be stressed that we see distributional sources of information as *contributing* towards an understanding of lexical processing difficulty; it is obviously not the whole story. A substantial amount of variability simply cannot be explained in terms of an expectation-building strategy, and consequently an (imperfect) model of this strategy would not be expected to account for a large portion of the total variability.

Of course, information available from the extra-linguistic context and general knowledge about the world also conceivably influence the processor's expectations about meaning. This is a limitation that does not diminish the value of the current approach, since our goal was essentially to provide an existence proof for the usefulness of the information inherent in the linguistic environment for understanding lexical processing effort.

In Chapter 7, the ICD model was revealed to be deficient in at least one important respect: it makes no provision for the role of syntactic structure. Morris (1994) demonstrated that syntactic differences modulated priming behaviour in her eye movement study. Although it is interesting how much variability can be accounted for without a syntactic component, it is clear that syntax is a highly influential factor.

One limitation that we have virtually ignored thus far is the fact that we do not expect our approach to provide accurate behavioural predictions for the class of function words. First, measuring the effort involved in processing function words is not without problems. Reaction times collected using word recognition tasks are subject to floor effects (Gordon & Caramazza, 1985), and eye movement measures are problematic simply because function words are frequently skipped, implying that they are either processed parafoveally during fixations made on the previous word, or else are not processed in a bottom-up fashion at all. Second, in English at least, function words seem to have a qualitatively different role from content words to play in prediction. Although function words are strong predictors of the syntactic category of an upcoming word, they do not seem to build expectations about its meaning. Third, as the object of prediction, function words are strongly predicted only when they occur as privileged components of a lexical unit, such as the phrasal verb *blurt out*. These predictions are clearly not contingent on expectations about meaning, rather, they are syntagmatic in nature. This type of information is certainly present in the linguistic environment; simple (local) co-occurrence statistics contain such information, which could be exploited by the language processor as a strategy complementary to the formation of expectations about meaning.

Finally, the information-oriented model does not make any predictions for the effort involved in processing rare (and in the extreme case, unknown) words. One reason is practical: statistics based on small samples are unreliable. Expanding the size of the corpus would yield more reliable co-occurrence statistics for those words defined as rare in the original corpus. However, if a subject has not had any experience with a given word (it is *unknown*), then no amount of corpus evidence for that word would be relevant for making behavioural predictions for that particular subject. Because we have put forward [I]CD as a model of the effort in recovering word meaning, we assume that the difficulty of recognising an unknown word presented in isolation reflects the failure of the processor to retrieve its meaning. Processing effort is therefore a combination of this failure and other, task-dependent processes. In the case that the unknown word is encountered in context, the context can provide clues that enable meaning to be inferred (*eg*. Carnine, Kameenui & Coyle, 1984; Fischer, 1994). We assume that this type of contextual inference involves higher-level cognitive processing than what is requires for word recognition, and consequently would not expect response time variability to be captured by a model such as ICD.

The [I]CD model could, in principle, model the lexical processing effort incurred by an individual for all words in his or her lexicon, if a complete environmental record was available, *ie.* if a corpus of all language exposure could be collected for that person. At present this is impossible; therefore we believe that refraining from making processing predictions for rare words, for a population of subjects, is reasonable given the inherent variability in the definition of 'rare' for a given member of this population.

8.5 Future directions

There are numerous possibilities for extending and elaborating upon the research programme begun in this thesis. Much work remains to be done in evaluating CD and ICD against experimental results. The more empirical data that can be satisfactorily explained in terms of measures derived from distributional information, the stronger the case for the role of the environment in explaining lexical processing effort.

Modelling word recognition: Extensions

The role of CD as a predictor of isolated word recognition behaviour has so far been restricted to the visual lexical decision task. The naming (or pronunciation) task is also widely held to be sensitive to semantic variables, even though the obtained results are often both quantitatively and qualitatively different from those of the lexical decision task. Successful modelling of naming latencies would reinforce CD as a psychologically plausible measure of lexical processing effort.

Word recognition is also a necessary component of tasks such as translation, which presumably involves semantic processing (*eg*. de Groot, 1992). Recent work by Tokowicz (1997) suggests that translation performance is affected by the number of translation equivalents a given word has. CD offers a plausible quantitative alternative to this variable, and can be calculated for both the source language and target language members of a translation pair.

Functional pressures on syntactic order

In the domain of syntactic processing, functional (or performance) considerations have been claimed to underlie cross-linguistic preferences for syntactic order (Hawkins, 1994). In Hawkins view, ordering conventions are motivated by 'least

effort' principles – real-time processing constraints. CD, as a measure of lexical processing effort, would also be expected to predict preferred orders in certain linguistic domains.

One domain which appears to be structured, though not according to syntactic conventions, is adjectival ordering. In English, a sequence of consecutive adjectives can premodify a noun, but these adjectives cannot appear in an arbitrary order; for example, *current global recession* is possible, but not *global current recession*. We propose that the relative ordering of premodifying adjectives in English is a monotonic function of the estimated effort involved in processing each adjective in the sequence. The simplest hypothesis derivable from such processing considerations is that the adjective which conveys the least information should be preferred in initial position, and the adjective with the highest CD value should be placed just before the noun.

Lexical selection in machine translation

A computational model of the psychological concept of contextual constraint could be readily put to good use for certain tasks in NLP. For example, the problem of target word selection in machine translation is a task that could benefit from estimates of contextual constraint. The general idea is that for cases where a source language word can be translated into more than one target language word, the preferred candidate will be more strongly constrained by the (target language) context. We have already carried out initial work on this problem using the Semantic Congruity measure developed in Chapter 4, with promising results (*cf.* McDonald, 1998b).

8.6 Final words

Has this thesis contributed towards an understanding of lexical processing effort? The answer is definitely "yes". Above all, we have provided evidence that environmental factors have an important role to play in explanations of lexical processing behaviour, and we have argued that these factors should be made prominent in theories of language processing. It is exposure to the linguistic environment, not the properties of the cognitive architectures and mechanisms underlying language processing, that we have claimed to be responsible for a significant portion of the measurable variability in lexical processing effort.

Bibliography

Altarriba, J., Kroll, J., Sholl, A. & Rayner, K. (1996). The influence of lexical and conceptual constraints on reading mixed-language sentences: Evidence from eye fixations and naming times. *Memory & Cognition*, *24*, 477-492.

Altmann, G. T. M. & Steedman, M. (1988) Interaction with context during human sentence processing. *Cognition*, *30*, 191-238.

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Baayen, R. H., Dijkstra, T. & Schreuder, R. (1997). Singulars and plurals in Dutch: Evidence for a parallel dual-route model. *Journal of Memory & Language*, *37*, 94-117.

Baayen, R. H., Piepenbrock, R. & van Rijn, H. (1993). *The CELEX lexical database (CD-ROM)*. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Baddeley, A. D. (1990). *Human memory: Theory and practice*. Hove: Erlbaum.

Balota, D. A. & Paul, S. T. (1996). Summation of activation: Evidence from multiple primes that converge and diverge within semantic memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 827-845.

Balota, D. A., Pollatsek, A. & Rayner, K. (1985). The interaction of contextual constraints and parafoveal visual information in reading. *Cognitive Psychology*, *17*, 364-390.

Bard, E. G., Shillcock, R. C. & Altmann, G. T. M. (1988). The recognition of words after their acoustic offsets in spontaneous speech: Effects of subsequent context. *Perception and Psychophysics*, *44*, 81-94.

Barsalou, L. W. (1982). Context-independent and context-dependent information in concepts. *Memory & Cognition*, *10*, 82-93.

Becker, C. A. (1980). Semantic context effects in visual word recognition. *Memory & Cognition*, *8*, 493-512.

Borowsky, R. & Masson, M. E. J. (1996). Semantic ambiguity effects in word identification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 63-85.

Bradley, D. C. (1983). *Computational distinctions of vocabulary type*. Bloomington, Indiana: Indiana University Linguistics Club.

Brodeur, D. A. & Lupker, S. J. (1994). Investigating the effects of multiple primes. *Psychological Research*, *57*, 1-14.

Bullinaria, J. & Huckle, C. C. (1998). Modelling lexical decision using corpus derived semantic representations in a connectionist network. In J. A. Bullinaria, D. Glasspool & G. Houghton (Eds.) *Proceedings of the 4th Neural Computation and Psychology Workshop, London, 9-11 April 1997* (pp. 213-226). London: Springer-Verlag.

Burgess, C. & Lund, K. (in press). The dynamics of meaning in memory. In Dietrich & Markman (Eds.) *Cognitive dynamics: Conceptual change in humans and machines.*

Burnage, G. & Dunlop, D. (1992). Encoding the British National Corpus. In *Papers from the 13th International Conference on English Language Research on Computerised Corpora.*

Cann, R. (in press). Functional versus lexical: a cognitive dichotomy. In R. D. Borsley (Ed.) *The nature and function of syntactic categories (Syntax and Semantics 32).* London: Academic Press.

Caplan, D. & Waters, G. S. (1999). Verbal working memory and sentence comprehension. *Behavioral and Brain Sciences, 22,* 77-94.

Carnine, D., Kameenui, E. J. & Coyle, G. (1984). Utilisation of contextual information in determining the meaning of unfamiliar words. *Reading Research Quarterly, 19,* 188-204.

Carroll, J. B. (1970). An alternative to Juilland's Usage Coefficient for lexical frequencies and a proposal for a Standard Frequency Index (SFI). *Computer Studies in the Humanities and Verbal Behaviour, 3,* 61-65.

Chater, N. (1989). *Information and information processing.* Unpublished PhD dissertation, University of Edinburgh.

Chater, N. & Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in Cognitive Sciences, 3,* 57-65.

Chenery, H. J., Murdoch, B. E. & Ingram, J. C. (1996). An investigation of confrontation naming performance in Alzheimer's dementia as a function of disease severity. *Aphasiology, 10,* 423-441.

Chiarello, C., Burgess, C., Richards, L. & Pollock, A. (1990). Semantic and associative priming in the cerebral hemispheres: Some words do, some words don't ... sometimes, some places. *Brain & Language, 38,* 75-104.

Church, K. W. & Gale, W. A. (1991). A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams. *Computer Speech and Language, 5,* 19-54.

Church, K. W. & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics, 16,* 22-29.

Cohen, J. D., MacWhinney, B., Flatt, M. & Provost, J. (1993). PsyScope: a new graphic interactive environment for designing psychology experiments. *Behavior Research Methods, Instruments & Computers, 25,* 257-271.

Collins, E. M. & Loftus, E. F. (1975). A spreading activation theory of semantic processing. *Psychological Review, 82,* 407-428.

Coltheart, M. (1981). The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology, 33A,* 497-505.

Coltheart, M., Davelaar, E., Jonasson, J. T. & Besner, T. (1977). Access to the internal lexicon. In S. Dornic (Ed.) *Attention and performance VI* (pp. 535-555). Hillsdale, NJ: Erlbaum.

Cruse, D. A. (1986). *Lexical semantics.* Cambridge: Cambridge University Press.

Cutler, A. (1993). Phonological cues to open- and closed-class words in the processing of spoken sentences. *Journal of Psycholinguistic Research, 22,* 109-131.

Dagan, I., Lee, L. & Pereira, F. (1999). Similarity-based models of word cooccurrence probabilities. *Machine Learning, 34,* 43-69.

Dagan, I., Pereira, F. & Lee, L. (1994). Similarity-based estimation of word cooccurrence probabilities. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics* (pp. 272-278). Somerset, NJ: ACM.

Dahl, H. (1979). *Word frequencies of spoken American English.* Essex, CT: Verbatim.

de Groot, A. M. B. (1992). Determinants of word translation. *Journal of Experimental Psychology: Learning, Memory and Cognition, 5,* 1001-1018.

Dennis, S. (1994). *The integration of learning into models of human memory*. Unpublished PhD dissertation, University of Queensland.

Duffy, S. A., Henderson, J. M. & Morris, R. K. (1989). The semantic facilitation of lexical access during sentence processing. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *15*, 791-801.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, *19*, 61-74.

Durkin, K. & Manning, J. (1989). Polysemy and the subjective lexicon: Semantic relatedness and the salience of intraword senses. *Journal of Psycholinguistic Research*, *18*, 577-612.

Faust, M. (1998). Obtaining evidence of language comprehension from sentence priming. In M. Beeman & C. Chiarello (Eds.) *Right hemisphere language comprehension* (pp. 160-185). Mahwah, NJ: Erlbaum.

Finch, S. & Chater, N. (1992). Bootstrapping syntactic categories. In *Proceedings of the 14th Annual Conference of the Cognitive Science Society* (pp. 820-825). Hillsdale, NJ: Erlbaum.

Fischer, U. (1994). Learning words from context and dictionaries: an experimental approach. *Applied Psycholinguistics*, *15*, 551-574.

Fischler, I. (1977). Semantic facilitation without association in a lexical decision task. *Memory & Cognition*, *5*, 335-339.

Forster, K. I. (1979). Levels of processing and the structure of the language processor. In W. E. Cooper & E. Walker (Eds.) *Sentence processing: Psycholinguistic studies presented to Merrill Garrett* (pp. 27-85). Hillsdale, NJ: Erlbaum.

Gallant, S. I. (1991). A practical approach for representing context and for performing word sense disambiguation using neural networks. *Neural Computation*, *3*, 293-309.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (1995). *Bayesian data analysis*. London: Chapman & Hall.

Gernsbacher, M. A. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General*, *113*, 256-281.

Goodman, N. (1972). Seven strictures on similarity. In N. Goodman (Ed.) *Problems and projects* (pp. 437-447). New York: Bobbs-Merrill.

Gordon, B. & Caramazza, A. (1985). Lexical access and frequency sensitivity – frequency saturation and open closed class equivalence. *Cognition*, *21*, 95-115.

Grefenstette, G. (1994). *Explorations in automatic thesaurus discovery*. Boston: Kluwer.

Griffin, Z. M. & Bock, K. (1998). Constraint, word frequency and the relationship between lexical processing levels in spoken word production. *Journal of Memory & Language*, *38*, 313-338.

Harris, Z. S. (1954). Distributional structure. *Word*, *10*, 140-162.

Hawkins, J. A. (1994). *A performance theory of order and constituency*. Cambridge: Cambridge University Press.

Hodgson, J. M. (1991). Informational constraints on pre-lexical priming. *Language and Cognitive Processes*, *6*, 169-205.

Hofland, K. & Johansson, S. (1982). *Word frequencies in British and American usage*. Harlow, UK: Longman.

Huckle, C. (1996). *Unsupervised categorisation of word meanings using statistical and neural network models.* Unpublished PhD dissertation, University of Edinburgh.

Hyona, J. (1993). Effects of thematic and lexical priming on readers' eye movements. *Scandinavian Journal of Psychology*, *34*, 293-304.

James, C. T. (1975). The role of semantic information in lexical decisions. *Journal of Experimental Psychology: Human Perception & Performance*, *1*, 130-136.

Jastrzembski, J. E. (1981). Multiple meanings, number of related meanings, frequency of occurrence, and the lexicon. *Cognitive Psychology, 13,* 278-305.

Jelinek, F., Mercer, R. L. & Roukos, S. (1992). Principles of lexical language modelling for speech recognition. In S. Furui and M. M. Sondhi (Eds.) *Advances in speech signal processing.* Maral Dekku.

Joordens, S. & Becker, S. (1997). The long and short of semantic priming effects in lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 1083-1105.

Jorgensen, J. C. (1990). The psychological reality of word senses. *Journal of Psycholinguistic Research, 19,* 167-190.

Just, M. A. & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 87,* 329-354.

Kaplan, E., Goodglass, H. & Weintraub, S. (1983). *The Boston naming test.* Philadelphia, PA: Lea & Febiger.

Katz, S. M. (1996). Distribution of context words and phrases in text and language modelling. *Journal of Natural Language Engineering, 2,* 15-59

Kirchner, H. S., Webb, W. G. & Kelly, M. P. (1984). The naming disorder of dementia. *Neuropsychologia, 22,* 23-30.

Kiss, G. R., Armstrong, C., Milroy, R. & Piper, J. (1973). An associative thesaurus of English and its computer analysis. In A. J. Aitkin, R. W. Bailey & N. Hamilton-Smith (Eds.) *The computer and literary studies.* Edinburgh: Edinburgh University Press.

Kliegl, R., Olson, R. K. & Davidson, B. J. (1982). Regression analyses as a tool for studying reading processes: Comments on Just and Carpenter's eye fixation theory. *Memory & Cognition, 10,* 287-296.

Kozima, H. & Ito, A. (1995). Context-sensitive measurement of word distance by adaptive scaling of a semantic space. In *Proceedings of Recent Advances in Natural Language Processing* (pp. 161-168). Tzigov Chark, Bulgaria.

Landauer, T. K. & Dumais, S. T. (1997). A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104,* 211-240.

Landauer, T. K., Foltz, P. W. & Laham, D. (1998). Introduction to Latent Semantic Analysis, *Discourse Processes, 25,* 259-284.

Lapata, M., McDonald, S. & Keller, F. (1999). Determinants of adjective-noun plausibility. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 30-36). Bergen, Norway.

Lauer, M. (1995). *Designing statistical language learners: Experiments on noun compounds.* Unpublished PhD dissertation, Macquarie University.

Levy, J. P, Bullinaria J. A. & Patel, M. (1997). Using co-occurrence statistics for modelling language processes. Poster presented at the *3rd Conference on Architectures and Mechanisms for Language Processing (AMLaP'97),* Edinburgh, Scotland. September 11-13, 1997.

Lorch, R. F. & Myers, J. L. (1990). Regression analyses of repeated measures data in cognitive research. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 149-157.

Lovelace, E. A. (1988). On using norms for low-frequency words. *Bulletin of the Psychomonic Society,* 26, 410-412.

Lowe, W. (1997) Meaning and the mental lexicon. In *Proceedings of the 15th International Joint Conference on Artificial Intelligence* (pp. 1092-1097). San Francisco: Morgan Kaufmann.

Lund, K. & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers, 28,* 203-208.

Lund, K., Burgess, C. & Atchley, R. (1995). Semantic and associative priming in high-dimensional semantic space. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society* (pp. 660-665). Mahwah, NJ: Erlbaum.

Lund, K., Burgess, C. & Audet, C. (1996). Dissociating semantic and associative word relationships using high-dimensional semantic space. In *Proceedings of the 18th Annual Conference of the Cognitive Science Society* (pp. 603-608). Mahwah, NJ: Erlbaum.

Mahalanobis, P. C. (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Science of India, 12,* 49-55.

Marr, D. (1982). *Vision*. San Francisco: W. H. Freeman.

Marslen-Wilson, W. D. & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition, 8,* 1-71.

Masson, M. E. J. (1986) Comprehension of rapidly presented sentences – the mind is quicker than the eye. *Journal of Memory & Language, 25,* 588-604.

McDonald, S. (1999). Capturing semantic influences on eye movements in reading. Poster presented at the *5th Conference on Architectures and Mechanisms for Language Processing (AMLaP'99),* University of Edinburgh, Edinburgh, Scotland. September 23-25, 1999.

McDonald, S. (1998a). Corpus-derived correlates of semantic similarity. To appear in *Proceedings of the XVI Congreso Nacional de la Asociación Española de Lingüística Aplicada,* Logroño, Spain. April 22-25, 1998.

McDonald, S. (1998b). Target word selection as proximity in semantic space. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Compuational Linguistics* (pp. 1496-1498). San Francisco: Morgan Kaufmann.

McDonald, S. (1997). Exploring the validity of corpus-derived measures of semantic similarity. Paper presented at the *9th Annual CCS/HCRC Postgraduate Conference,* University of Edinburgh. June 18-19, 1997. (`http://www.iccs.informatics.ed.ac.uk/~scottm/reliability_paper.w5.1.ps.gz`)

McDonald, S. & Lowe, W. (1998). Modelling functional priming and the associative boost. In *Proceedings of the 20th Annual Conference of the Cognitive Science Society* (pp. 667-680). Mahwah, NJ: Erlbaum.

McKoon, G. & Ratcliff, R. (1992). Spreading activation versus compound cue accounts of priming: Mediated priming revisited. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 1155-1172.

McNamara, T. P. & Miller, D. L. (1989). Attributes of theories of meaning. *Psychological Bulletin, 106,* 377-394.

McRae, K. & Boisvert, S. (1996). The importance of automatic semantic relatedness priming for distributed models of word meaning. In*Proceedings of the 18th Annual Conference of the Cognitive Science Society* (pp. 278-283). Mahwah, NJ: Erlbaum.

Medin, D. L., Goldstone, R. L. & Getner, D. (1993). Respects for similarity. *Psychological Review, 100,* 254-278.

Meyer, D. & Schvaneveldt, R. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology, 90,* 227-234.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D. & Miller, K. (1990). Introduction to WordNet: an online lexical database. *International Journal of Lexicography, 3,* 235-244.

Miller, G. A. & Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes, 6,* 1-28.

Miller, G. A., Heise. G. & Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology, 41,* 329-335.

Monsell, S. (1991). The nature and locus of word frequency effects in reading. In D. Besner & G. W. Humphreys (Eds.) *Basic processes in reading: Visual word recognition*. Hillsdale, NJ: Erlbaum.

Morris, R. K. (1994). Lexical and message-level sentence context effects on fixation times in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 92-103.

Morrison, C. & Ellis, A. (1995). Roles of word frequency and age of acquisition in word naming and lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 116-133.

Moss, H. E., Hare, M. L., R. K., Day, P. & Tyler, L. K. (1994). A distributed memory model of the associative boost in semantic priming. *Connection Science*, *6*, 413-427.

Moss, H. E. & Marslen-Wilson, W. D. (1993). Access to word meanings during spoken language comprehension: Effects of sentential semantic context. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 1254-1276.

Moss, H. E., Ostrin, R. K., Tyler, L. K., & Marslen-Wilson, W. D. (1995). Accessing different types of lexical semantic information: Evidence from priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 863-883.

Nebes, R. D. (1989). Semantic memory in Alzheimer's Disease. *Psychological Bulletin*, *106*, 355-376.

Nebes, R. D. & Brady, C. B. (1991). The effect of contextual constraint on semantic judgments by Alzheimer patients. *Cortex*, *27*, 237-246.

Nebes, R. D., Boller, F. & Holland, A. (1986). Use of semantic context by patients with Alzheimer's Disease. *Psychology and Aging*, *1*, 261-269.

Neely, J. H. (1991). Semantic priming effects in visual word recognition: a selective review of current findings and theories. In D. Besner & G. W. Humphrey (Eds.) *Basic processes in reading: Visual word recognition* (pp. 264-336). Hillsdale, NJ: Erlbaum.

O'Seaghdha, P. G. (1989). The dependence of lexical relatedness effects on syntactic connectedness. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 73-87.

Oaksford, M. & Chater, N. (1998). An introduction to rational models of cognition. In M. Oaksford & N. Chater (Eds.) *Rational Models of Cognition* (pp. 1-18). Oxford: OUP.

Osgoode, C. E., Suci, G. J. & Tannenbaum, P. H. (1957). *The measurement of meaning*. Urbana, IL: University of Illinois Press.

Paivio, A. (1971). *Imagery and verbal processes*. New York: Holt, Rinehart & Winston.

Patel, M., Bullinaria, J. A. & Levy, J. P. (1998). Extracting semantic representations from large text corpora. In J. A. Bullinaria, D. Glasspool & G. Houghton (Eds.) *Proceedings of the 4th Neural Computation and Psychology Workshop, London, 9-11 April 1997* (pp. 199-212). London: Springer-Verlag.

Plaut, D. C. (1995). Semantic and associative priming in a distributed attractor network. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society* (pp. 37-42). Mahwah, NJ: Erlbaum.

Poesio M., Schulte im Walde, S. & Brew, C. (1997). Lexical clustering and definite description interpretation. In *Proceedings of the AAAI Spring Symposium on Learning for Discourse*. Stanford, CA.

Pulvermeyer, F., Lutzenberger, W. & Birbaumer, N. (1995). Electrocortical distinction of vocabulary types. *Electroencephalography and Clinical Neurophysiology*, *94*, 357-370.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*, 372-422.

Rayner, K. & Duffy, S.A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition*, *14*, 191-201.

Rayner, K., Sereno, S. C., Morris, R. K., Schmauder, A. R. & Clifton, C. (1989). Eye movements and on-line language comprehension processes. *Language and Cognitive Processes, 4*, 21-49.

Rayner, K. & Well, A. D. (1996). Effects of contextual constraint on eye movements in reading: a further examination. *Psychonomic Bulletin & Review, 3*, 504-509.

Redington, M., Chater, N. & Finch, S. (1998). Distributional information: a powerful cue for acquiring syntactic categories. *Cognitive Science, 22*, 425-469.

Redington, M., Chater, N., Huang, C., Chang, L., Finch, S. & Chen, K. (1995). The universality of simple distributional methods: Identifying syntactic categories in Mandarin Chinese. In *Proceedings of the 4th International Conference on the Cognitive Science of Natural Language Processing*. Dublin City University.

Reichle, E. D., Pollatsek, A., Fisher, D. L. & Rayner, K. (1998). Towards a model of eye movement control in reading. *Psychological Review, 105*, 125-157.

Resnik, P. S. (1993). *Selection and information: a class-based approach to lexical relationships*. Unpublished PhD dissertation, University of Pennsylvania.

Resnik, P. S. (1997). Selectional preference and sense disambiguation. *ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?* Washington, DC. April 4-5, 1997.

Rommetveit, R. (1968). *Words, meanings, and messages: Theory and experiments in psycholinguistics*. New York: Academic Press.

Rubenstein, H. & Goodenough, J. B. (1965). Contextual correlates of synonymy. *Computational Linguistics, 8*, 627-633.

Schustack, M. W., Ehrlich, S. F. & Rayner, K. (1987). Local and global sources of contextual facilitation in reading. *Journal of Memory & Language, 26*, 322-340.

Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics, 24*, 97-124.

Schütze, H. (1993). Word space. In S. J. Hanson, J. D. Cowan & C. L. Giles (Eds.) *Advances in Neural Information Processing Systems 5* (pp. 895-902). San Mateo, CA: Morgan Kaufmann.

Schütze, H. (1992). Dimensions of meaning. In *Supercomputing '92: Proceedings of the ACM/IEEE Conference* (pp. 787-796). Los Alamitos, CA: IEEE Computer Society Press.

Schwanenflugel, P. J. (1986). Completion norms for final words of sentences using a multiple production measure. *Behavior Research Methods, Instruments, & Computers, 18*, 363-371.

Schwanenflugel, P. J. & LaCount, K. L. (1988). Semantic relatedness and the scope of facilitation for upcoming words in sentences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 344-354.

Schwanenflugel, P. J. & Shoben, E. J. (1985). The influence of sentence constraint on the scope of facilitation for upcoming words. *Journal of Memory and Language, 24*, 232-252.

Schwanenflugel, P. J. & Shoben, E. J. (1983). Differential context effects in the comprehension of abstract and concrete verbal materials. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9*, 82-102.

Schwanenflugel, P. J., Harnishfeger, K. K. & Stowe, R. W. (1988). Context availability and lexical decisions for abstract and concrete words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 499-520.

Segui, J., Frauenfelder, U. H., Lainé, C. & Mehler, J. (1987). The word frequency effect for open- and closed-class items. *Cognitive Neuropsychology, 3*, 33-44.

Shelton, J. R. & Martin, R. C. (1992). How semantic is automatic semantic priming? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*, 1191-1210.

Shillcock, R. C. & Bard, E. G. (1993). Modularity and the processing of closed class words. In G. T. M. Altmann & R. C. Shillcock (Eds.) *Cognitive models of speech processing: the 2nd Sperlonga Meeting*. Erlbaum.

Siegel, S. & Castellan, N. J. (1988). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.

Simpson, G. B., Peterson, R. R., Casteel, M. A. and Burgess, C. (1989). Lexical and sentence context effects in word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 88-97.

Spence, D. P., & Owens, K. C. (1990). Lexical co-occurrence and association strength. *Journal of Psycholinguistic Research*, *19*, 317-330.

Steedman, M. (1989). Grammar, interpretation and processing from the lexicon. In W. Marslen-Wilson (Ed.) *Lexical representation and process* (pp. 463-504). Cambridge, Mass.: MIT Press.

Stolcke, A. (1994). *Bayesian learning of probabilistic language models*. Unpublished PhD dissertation, University of California at Berkeley.

Svartvik, J. & Quirk, R. (Eds.) (1980). A corpus of English conversation. *Lund Studies in English 56*. Lund: Lund University Press.

Swinney, D. A., Zurif, E. B. & Cutler, A. (1980). Effects of sentential stress and word class upon comprehension in Broca's aphasics. *Brain & Language*, *10*, 132-144.

Tabossi, P. (1988). Effects of context upon the immediate interpretation of unambiguous words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 153-162.

Taft, M. (1979). Recognition of affixed words and the word frequency effect. *Memory & Cognition*, *7*, 263-272.

Tanenhaus, M. K. & Lucas, M. M. (1987). Context effects in lexical processing. *Cognition*, *25*, 213-234.

Tokowicz, N. (1997). *Reevaluating concreteness effects in bilingual translation*. Unpublished Master's thesis, Department of Psychology, The Pennsylvania State University.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*, 327-352.

Van Berkum, J., Hagoort, P. & Brown, C. (1998). Semantic integration in single sentences and stories: Evidence from the N400. Poster presented at the *4th Conference on Architectures and Mechanisms for Language Processing (AMLaP'98)*, University of Freiburg, Freiburg, Germany. September 24-26, 1998.

Van Petten, C., Coulson, S., Rubin, S., Plante, E. & Parks, M. (1999). Time course of word identification and semantic integration in spoken language. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *25*, 394-417.

Vu, H., Kellas, G. & Paul, S. T. (1998). Sources of sentence constraint on lexical ambiguity resolution. *Memory & Cognition*, *26*, 979-1001.

Williams, J. N. (1988). Constraints upon semantic activation during sentence comprehension. *Language and Cognitive Processes*, *3*, 165-206.

Williams, J. N. (1996). Is automatic priming semantic? *European Journal of Cognitive Psychology*, *8*, 113-161.

Zipf, G. K. (1935). *The psycho-biology of language*. Boston: Houghton Mifflin.

Appendices

Appendix A

In Chapter 2, we attempted to provide psychological motivations for most of the parameter settings used in the basic representational model investigated in Chapters 3 and 4. It is the purpose of this appendix to demonstrate further support for these settings using purely empirical criteria.

In some sense, the modelling results reported in Chapters 3 and 4 are reliant on the model parameters chosen being near-optimal, and it would be useful to examine their influence more closely, be exploring the parameter space around the chosen settings. In this section we report the results of a series of parameter manipulations, using the semantic similarity ratings elicited for 30 word pairs in Experiment 2A, Chapter 3, as the evaluation standard. Specifically, the coefficient of determination (r^2), the amount of human judgement variance accounted for by semantic distance, is tracked while varying the parameter settings. This evaluation method assesses the degree to which the similarity relationships between words are captured by the similarity between co-occurrence vector representations for those words. It is important to bear in mind that r^2 is substantially dependent on the sample, and as a result a particular parameter exploration may look quite different when carried out for a second sample of word pairs. In addition, the parameter settings clearly interact; for instance, the optimal window size for this sample also depends on the semantic distance measure and the dimensionality of the space.

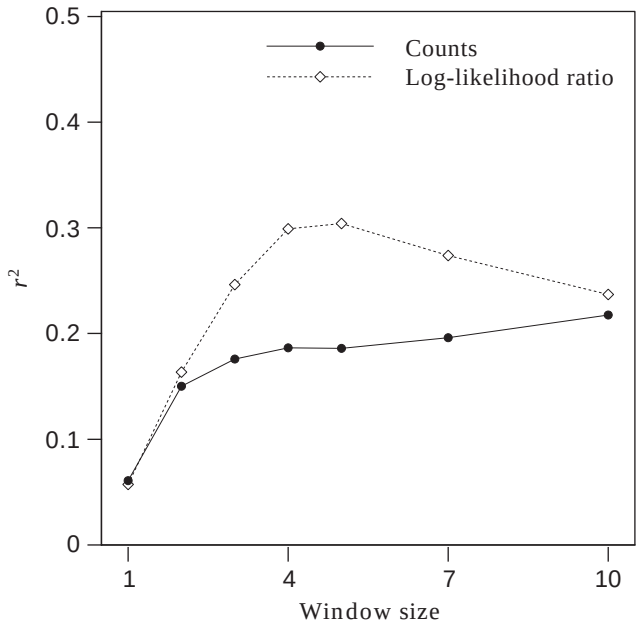

Figure A-1. The effects of window size and how co-occurrence is encoded.

In the sections below, we graphically present the effect of varying one or two model parameters on r^2, while holding the remaining parameters constant.

Window size and component values

Figure A-1 displays the results of varying the window size on the amount of similarity rating variance accounted for by semantic distance, for two approaches to encoding the co-occurrence relationship between target and context words. The set of 446 content words was used to define the dimensions of the semantic space, and the cosine of the angle between vectors served as the semantic distance measure.

The results of simultaneously varying window direction and size are indicated in Figure A-2. Other model parameters were held constant (see above), and vector components encoded the log-likelihood ratio statistic. Note that a *Before and after* window size of three means three words to either side of the target (for an effective window size of six). Contrary to Levy *et al.*'s (1997) findings, recording co-occurrence using an *After only* window resulted in slightly better performance than a *Before and after* window.

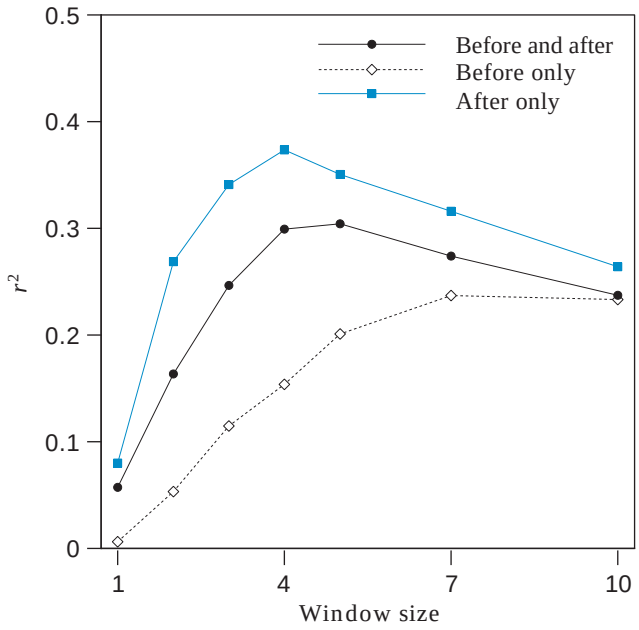

Figure A-2. The effects of window direction and size on goodness-of-fit.

Selection of context words

Figure A-3 displays the results of manipulating the context word set used to define the dimensions of the representational space. Using a window size of ±3 words, the log-likelihood ratio to encode co-occurrence, and the cosine of the angle between vectors as the semantic distance measure, both the number and selection of context words were manipulated. The *Function & content* and the *Content only* sets were created using the topmost *k* items from the BNC-spoken lexeme frequency list; the *Content only* set excludes function words (see Appendix B).

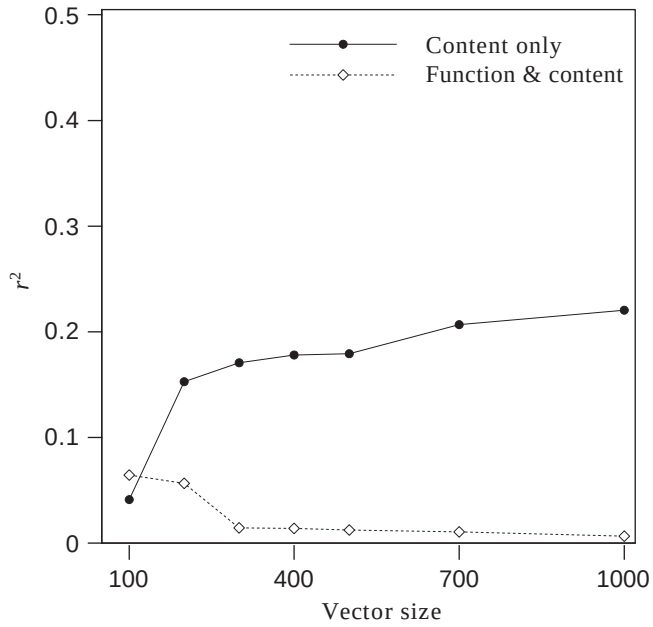

Figure A-3. The effects of vector size and the function-content word distinction.

Appendix B

Membership of the set of function words was determined according to the method described in section 2.3.3. The 319 function words listed on the following page were excluded from consideration as vector components, and were filtered from sentence contexts in all computational simulations involving context (*eg.* Experiments 6, 7 and 8 in Chapter 4).

’d
’em
’neath
’s
’tween
’twixt
’un
a
a la
abaft
aboard
about
above
according as
according to
across
across from
after
against
agin
albeit
all
along
alongside
although
amid
amidst
among
amongst
an
and
anent
any
anyone
anything
apropos of
around
as
as far as
as for
as from
as if
as long as
as of
as opposed to
as regards
as soon as
as though
as to
as well as
ascribable to
aside from
astride
at
athwart
atop
aught
barring
be
because
because of
before
behind
below
beside
between
betwixt
beyond
both
both and
but
but for
but that
but then
but then again
by
by virtue of
can
certain
circa
concerning
considering
contrary to
cos
deep in
despite
do
due to
during
each
each other
either or
ere
even if
even though
every
everybody
everything
ex libris
except
excepting
excluding
exclusive of
fewer
fewest
for
forasmuch as
former
from
further
further to
half
hard on
hard upon
have
he
her
hers
herself
him
himself
his
how
howsoever
i
idem
if
in
in as far as
in case
in memoriam
in so far as
inasmuch as
including
inside
inside of
insomuch as
insomuch that
instead of
into
irrespective of
it
its
itself
last
less
lest
let’s
like
many
may
me
mine
minus
more
most
must
my
myself
near
neither
neither nor
next
next to
no
no one
nobody
none
nor
not
nothing
notwithstanding
now that
o’
o’er
of
on
one another
oneself
onto
opposite to
or
other
other than
ought
our
ours
ourselves
out of
outside
outside of
over
owing to
own
per
plenty
plus
preliminary to
preparatory to
previous to
prior to
provided
provided that
providing
providing that
pursuant to
qua
re
regarding
regardless of
respecting
round
same
sans
save
seeing
seeing as
seeing as how
seeing that
several
shall
she
short of
since
so that
some
somebody
someone
something
subject to
such
such and such
suchlike
summat
supposing
than
that
the
thee
their
theirs
them
themselves
these
they
thine
this
those
thou
though
through
throughout
thru
thy
thyself
till
to
together with
toward
towards
under
unless
unlike
until
unto
up to
upon
upward of
upwards of
us
versus
via
vis-a-vis
we
what
whate'er
whatever
when
whenever
where
whereas
whereat
whereby
wherefore
wherein
whereon
wheresoever
whereunto
whereupon
wherever
whether
which
whichever
whichsoever
while
whilst
who
whoever
whom
whomever
whomsoever
whose
whoso
whosoever
why
will
with
within
without
worth
ya
ye
yet
you
you-all
your
yours
yourself
yourselves

Appendix C

Materials from Experiment 5, Chapter 4, with relevance-weighted Contextual Similarity measurements.

	Context	Ambiguous Word (AW)	Related Target (R)	Unrelated Control (C)	Weighted (AW↔R)	Similarity (AW↔C)
1.	(A) boat, dock	vessel	ship	beat	0.682	0.407
	(I) empty, jug	vessel	ship	beat	0.205	0.318
2.	(A) different, distribution	uniform	same	nice	0.731	0.848
	(I) soldier, army	uniform	same	nice	0.483	0.589
3.	(A) plate, bean	tin	lid	pet	0.743	0.150
	(I) mine, lead	tin	lid	pet	0.467	0.095
4.	(A) name, page	title	book	case	0.836	0.359
	(I) lord, sir	title	book	case	0.414	0.427
5.	(A) paper, open	tear	rip	toe	0.266	0.324
	(I) cry, drop	tear	rip	toe	0.275	0.435
6.	(A) door, straw	stable	horse	equal	0.276	0.108
	(I) balanced, diet	stable	horse	equal	0.362	0.155
7.	(A) moon, rocket	space	time	very	0.517	0.936
	(I) room, air	space	time	very	0.621	0.764
8.	(A) tap, wash	sink	kitchen	college	0.594	0.460
	(I) ship, drown	sink	kitchen	college	0.708	0.567
9.	(A) time, hand	second	minute	happen	0.511	0.147
	(I) first, third	second	minute	happen	0.400	0.412
10.	(A) factory, technology	production	line	view	0.469	0.081
	(I) play, film	production	line	view	0.230	0.020
11.	(A) drink, sherry	port	wine	king	0.603	0.200
	(I) harbour, ship	port	wine	king	0.406	0.361
12.	(A) bath, sink	plug	hole	race	0.574	0.322
	(I) spark, socket	plug	hole	race	0.385	0.175
13.	(A) time, work	play	game	wish	0.791	0.109
	(I) theatre, act	play	game	wish	0.917	0.195
14.	(A) hospital, bed	patient	doctor	father	0.466	0.109
	(I) calm, kind	patient	doctor	father	0.311	0.089
15.	(A) apple, fruit	orange	lemon	aside	0.547	0.179
	(I) colour, red	orange	lemon	aside	0.359	0.068
16.	(A) strange, peculiar	odd	queer	chick	0.248	0.208
	(I) even, one	odd	queer	chick	0.141	0.158
17.	(A) conceal, shame	hide	seek	glad	0.180	0.370
	(I) cow, leather	hide	seek	glad	0.128	0.278
18.	(A) beer, milk	drink	water	stuff	0.855	0.747
	(I) eat, glass	drink	water	stuff	0.837	0.709
19.	(A) scotland, america	country	england	average	0.371	0.063
	(I) town, house	country	england	average	0.247	0.049
20.	(A) ice, warm	cold	hot	dry	0.917	0.524
	(I) nose, sneeze	cold	hot	dry	0.924	0.502

Note: A=Appropriate context, I=Inappropriate context (in relation to the 'meaning' of the ambiguous item that the Related target word stands for).

Appendix D

Listed here are the materials from Experiment 6, Chapter 4. Sentence contexts are from Griffin and Bock (1998). Target words and control words are listed in their uninflected (citation) forms. The critical context words used to estimate semantic congruity are indicated in *italics*.

	Target	Control	Context
1.	car	job	George *taught* his *son* to *drive* a ___.
2.	ear	hat	He couldn't *hear well* because of the *infection* in his *left* ___.
3.	hand	plan	*everyone* was *shocked* when Mark was *willing* to *shake* his *enemy*'s ___.
4.	leg	guy	The skier *fell* and *broke* his ___.
5.	window	nobody	To *get* some *cool air* in the *apartment*, they *put* the *fan* in the ___.
6.	brain	stamp	He was *afraid* that *drugs* would *damage* his ___.
7.	bed	war	Bob was *tired* so he *went* to ___.
8.	bomb	agent	The *plane exploded* because of a *hidden* ___.
9.	dress	range	The *bridesmaid wore* an *ugly* ___.
10.	clock	apple	They didn't *know* what *time* it was because they couldn't *find* a ___.
11.	door	shop	*Always knock* before you *open* my ___.
12.	nose	loss	Vic *sneezed* and *blew* his ___.
13.	arm	gas	The pitcher was *unable* to *throw* the ball because of his *broken* ___.
14.	glass	fight	She *poured* the *lemonade* into a *tall* ___.
15.	baby	hair	They *bought* a *crib* for the ___.
16.	book	idea	The *author signed* a *copy* of her *new* ___.
17.	pencil	threat	To *fill* in the *bubble sheet* the *student needed* a *sharp* ___.
18.	bottle	debate	There was *glass* all over the sidewalk from a *broken* ___.
19.	star	spot	The *hopeful girl wished* upon a ___.
20.	plant	score	No one *remembered* to *water* the ___.
21.	foot	page	The *clumsy man stepped* on her ___.
22.	eye	cup	She *put* a *contact lens* in her ___.
23.	key	bet	He couldn't *unlock* the *door* without the *right* ___.
24.	knife	alarm	He *stabbed* the *man* with a *sharp* ___.
25.	ring	room	He *bought* his *girlfriend* an *engagement* ___.
26.	box	tea	When his *new computer finally arrived*, he *ripped open* the ___.
27.	house	money	They *moved* into a *new* ___.
28.	gun	row	The *bank robber aimed* at the *security officer* and *fired* the ___.
29.	axe	fur	He *chopped down* the *tree* with an ___.
30.	hook	text	The *fisherman attached* the *worm* to the ___.
31.	lock	link	The *bike* was *protected* from *theft* by an *expensive* ___.
32.	match	sheet	He *lit* the *candle* with *one* ___.
33.	whistle	carrier	The *referee stopped* the *game* by *blowing* his ___.

34.	broom	chick	Bob *swept* the *floor* with the ___.
35.	bone	ward	While Suzy was *eating chicken*, she *choked* on a ___.
36.	badge	coffin	To *prove* he was a *police officer*, he *showed* the *woman* his ___.
37.	drum	cure	The *people marched* to the *beat* of a *loud* ___.
38.	crown	theme	On *top* of his *head*, the *king wore* an *extremely expensive* ___.
39.	dog	kid	The *little puppy grew up* to be a *huge* ___.
40.	nail	fuel	The *wooden board* splintered when the *carpenter tried* to *insert* the ___.
41.	owl	mop	The *campers* were *frightened* by the *hoot* of an ___.
42.	ghost	shade	The *castle* was *haunted* by a *frightening* ___.
43.	button	survey	His *coat* was *open* because it was *missing* a ___.
44.	bee	cot	The *boy* was *stung* by the ___.
45.	scarf	lager	To *protect* his *neck* from the *cold* he *wore* a *long* ___.
46.	purse	pride	She *kept lipstick* and a *compact* in her ___.
47.	frog	arch	The *tadpole grew up* to be a *big* ___.
48.	egg	tie	The *hen laid* some ___.
49.	cow	pie	The *farmer milked* the ___.
50.	bat	mat	The *baseball player swung* the ___.
51.	knot	raid	The *sailor tied* the *rope* with a *complicated* ___.
52.	lion	cord	The *king* of the *jungle* is the ___.
53.	bowl	neck	She *ladled* the *soup* into her ___.
54.	heart	death	The valentine's *day card* was *shaped* like a ___.
55.	comb	rage	He *parted* his *hair* with a ___.

Appendix E

Materials from Experiment 7, Chapter 4 are listed in this Appendix. Sentence fragments and target words are taken from Schwanenflugel (1986). Note that targets and control words are listed in their uninflected forms. Critical context words are indicated in *italics.*

	Target	Control	Context
1.	diet	rose	The overweight *man went* on a ___.
2.	content	pudding	The *postman opened* the *package* to *inspect* its ___.
3.	stamp	clock	He *mailed* the *letter* without a ___.
4.	party	level	The *dispute* was *settled* by the *third* ___.
5.	oven	mail	The *woman took* the *warm cake* out of the ___.
6.	election	occasion	He *campaigned so* he would *win* the ___.
7.	fence	blank	To *keep animals* out of the *garden*, he *put up* a ___.
8.	mountain	platform	The hikers *slowly climbed up* the ___.
9.	kiss	bowl	My *uncle gave* my *mother* a *big* ___.
10.	window	police	Jill *looked back* through the *open* ___.
11.	locker	knight	John *kept* his *gym clothes* in a ___.
12.	bath	chip	The *tired mother gave* the *dirty child* a ___.
13.	beard	flour	The *old man* had a *long gray* ___.
14.	plate	fruit	He *scraped* the *cold food* from his ___.
15.	rain	data	The *picnic* was *spoiled* because of the ___.
16.	piano	slate	Harriet *sang* while my *brother played* the ___.
17.	letter	mother	While *away*, James *sent home* a ___.
18.	boss	path	The *worker* was *criticized* by his ___.
19.	fight	model	The *friends* were not *talking* because they had a ___.
20.	beard	stool	The *man decided* to *shave* his ___.
21.	tenant	impact	The *landlord* was *faced* with a *strike* by the ___.
22.	driveway	mobility	Our *new green car blocked* the ___.
23.	day	way	Most *students prefer* to *work* during the ___.
24.	love	shop	The *old man* and *woman married* for ___.
25.	table	piece	The *cup* was *placed* on the ___.
26.	beach	asset	On a *hot summer day* many *people go* to the ___.
27.	publicity	household	The *trial received a lot* of ___.
28.	illness	stomach	The *woman died* after a prolonged ___.
29.	work	year	*Too* many *men* are out of ___.
30.	ability	railway	The *difficult task* was beyond his ___.
31.	breakdown	workplace	*Too much stress* can *cause nervous* ___.
32.	direction	breakfast	The *car* in *front suddenly changed* ___.
33.	floor	space	Some of the *ashes dropped* on the ___.
34.	bag	bar	The *shopper carried home* several clumsy ___.
35.	meal	mess	Some *people* have *never* had a *square* ___.
36.	zoo	dam	The *wild animals* were *seen* in the ___.
37.	inflation	migration	More *money buys few products* during *times* of ___.
38.	mansion	amenity	The *wealthy businessman lived* in a ___.
39.	swimming	entrance	*Calm seas* are *always good* for ___.

40.	race	song	Jack *bet* all he had on the *last* ___.
41.	cave	lily	A *large stone blocked* the *entrance* to the ___.
42.	milk	boot	John *poured* himself a *glass* of ___.
43.	studio	autumn	The *artist painted* the masterpiece in his ___.
44.	law	air	He was *punished* for *breaking* the ___.
45.	depression	assistance	*Economic indicators predicted* a *future* ___.
46.	clothes	element	In *preparation* for the *trip* the *children packed* ___.
47.	scene	shirt	The wooded *lake made* a *pretty* ___.
48.	rate	game	The *drinking age* was *raised* because of the *accident* ___.
49.	fear	crap	Because he hadn't *studied*, Tom *faced* the *exam* with ___.
50.	movie	shade	We had to *wait* in *line* at the ___.
51.	arrest	boiler	*New clues led* to the *criminal's* ___.
52.	rhythm	pallet	*Dancing requires* a *sense* of ___.
53.	cemetery	syllabus	The *woman* was *laid* to *rest* in the ___.
54.	shelf	salad	Helen *reached up* to *dust* the ___.
55.	fire	view	The *woman asked* her *husband* to *put out* the ___.

Appendix F

Stimuli used for parameter optimisation (n=70) and their mean visual lexical decision response times (averaged over 18 subjects). Items are sorted by their contextual distinctiveness (CD) score.

		RT (msec)	
Word	CD (bits)	M	SD
scarf	2.396	556	72
worthy	2.306	516	100
shade	1.405	552	102
deserve	1.256	507	98
coffee	1.198	490	79
penny	1.132	497	87
station	1.033	494	77
willing	1.012	509	98
peace	0.978	510	89
wire	0.955	503	93
pocket	0.902	490	88
stroke	0.889	500	85
roof	0.881	483	80
green	0.867	505	95
large	0.865	516	87
blue	0.834	460	82
winter	0.833	479	103
door	0.811	483	89
private	0.802	449	85
railway	0.787	476	67
clothes	0.760	472	64
smell	0.757	459	78
hall	0.734	490	91
term	0.729	536	97
main	0.695	473	100
chicken	0.695	463	96
army	0.692	498	91
corner	0.662	471	77
morning	0.637	476	80
memory	0.626	472	81
summer	0.608	480	66
fault	0.604	492	80
hotel	0.597	483	57
floor	0.590	484	95
iron	0.587	514	104
accept	0.577	459	95
garden	0.561	481	76
mother	0.560	497	95
collect	0.530	495	94

chance	0.519	466	68
heart	0.513	479	80
office	0.507	523	99
write	0.486	453	98
paper	0.483	471	95
face	0.481	479	85
save	0.467	446	67
full	0.466	496	75
wife	0.462	491	103
away	0.461	489	98
turn	0.460	464	98
woman	0.458	481	75
hand	0.446	459	83
head	0.428	461	87
play	0.424	492	80
school	0.422	455	96
never	0.404	458	90
stage	0.392	495	73
machine	0.389	517	96
company	0.381	490	96
surely	0.353	509	91
wrong	0.342	491	87
sorry	0.325	464	94
back	0.313	477	76
mind	0.296	486	99
good	0.292	480	101
need	0.214	487	78
tell	0.207	512	84
look	0.176	487	79
some	0.168	471	92
right	0.152	458	77

Appendix G

Stimuli used in Experiment 9, Chapter 5 (*n*=53), and their visual lexical decision response times (averaged over 18 subjects). Items are sorted by CD.

		RT (msec)	
Word	CD (bits)	M	SD
comic	1.886	525	116
elbow	1.662	520	115
forty	1.480	565	119
rugby	1.421	569	138
twenty	1.297	510	114
lounge	1.241	504	126
refer	1.189	576	121
button	1.174	498	106
four	1.159	536	124
lord	1.132	513	139
century	1.096	543	123
piano	1.052	496	132
black	0.986	551	121
three	0.933	484	131
neck	0.931	562	127
effort	0.911	485	137
hill	0.869	500	116
stone	0.844	477	115
grand	0.766	541	107
death	0.751	504	105
street	0.744	499	109
theatre	0.737	524	115
season	0.712	500	118
lift	0.712	507	117
young	0.678	523	115
general	0.678	536	122
wind	0.630	537	130
second	0.600	523	127
remind	0.591	497	105
side	0.552	501	139
ready	0.512	513	118
lead	0.496	528	119
room	0.479	484	115
kind	0.472	465	111
enough	0.423	511	107
ever	0.418	545	119
down	0.377	511	110
bill	0.376	504	99
rather	0.296	510	108
exactly	0.291	556	129
call	0.277	499	142

come	0.256	496	112
moment	0.184	523	131
time	0.173	501	136
only	0.160	527	108
just	0.110	505	125
think	0.091	500	138

Appendix H

This appendix lists the word and nonword materials used in Experiment 10, Chapter 5. Below are the critical matched stimuli (*n*=40), including their visual lexical decision response times (averaged across 24 subjects), and five other lexical properties. Items are sorted by corpus frequency (lnLF).

Word		RT (msecs)			CD (bits)		lnLF		AoA		Familiarity		N	
H	L	H	L	Diff	H	L	H	L	H	L	H	L	H	L
pound	money	495	499	-4	1.696	0.542	9.143	8.795	308	247	618	631	8	3
council	problem	547	521	26	1.503	0.341	8.492	8.679	464	367	508	596	1	1
member	matter	480	538	-58	0.898	0.338	8.072	8.028	392	411	573	563	1	12
union	drink	498	501	-3	1.453	0.711	7.496	7.525	503	211	595	628	2	5
page	type	532	521	11	1.130	0.496	7.485	7.509	267	383	603	567	12	5
health	market	478	496	-18	1.515	0.615	7.380	7.413	400	328	577	518	3	3
income	normal	555	497	58	1.049	0.443	6.783	6.820	506	375	521	602	1	2
range	guess	541	527	14	0.969	0.338	6.568	6.596	436	292	515	585	3	2
safety	animal	510	472	38	1.617	0.584	6.553	6.532	339	222	556	620	2	1
relief	agency	569	565	4	2.026	1.092	5.976	5.979	443	553	551	420	2	1
primary	article	527	539	-12	1.803	1.308	5.927	5.932	297	406	497	533	2	1
verse	ideal	563	515	48	1.647	0.831	5.638	5.663	351	461	483	521	5	2
sin	dot	591	623	-32	1.727	1.088	5.576	5.572	400	219	501	524	20	17
justice	content	518	522	-4	1.657	0.872	5.565	5.580	500	389	522	553	1	6
deputy	mirror	538	484	54	1.818	1.015	5.545	5.549	433	258	462	593	2	1
failure	scratch	533	580	-47	1.679	1.005	5.323	5.303	439	269	542	553	1	1
diamond	dispute	526	549	-23	2.400	1.427	5.170	5.170	339	522	512	520	1	1
burden	reward	572	511	61	2.094	1.331	4.836	4.820	474	372	446	525	1	4
cube	flag	564	523	41	1.904	1.293	4.771	4.754	383	258	502	545	5	10
ocean	organ	554	485	69	2.057	1.323	4.727	4.719	317	356	526	510	1	1
inquiry	antique	546	546	0	2.438	1.789	4.575	4.575	483	439	485	484	2	1
trunk	wreck	515	571	-56	2.282	1.589	4.419	4.419	328	369	485	516	3	3
ballot	cellar	656	497	159	2.375	1.665	4.331	4.344	539	361	453	467	2	2
banker	pillow	514	524	-10	3.288	1.628	4.290	4.277	392	217	524	602	13	3
frog	arch	540	539	1	2.338	1.560	4.277	4.277	258	367	507	483	4	2
clash	blade	533	523	10	1.965	1.287	4.263	4.277	422	344	488	517	7	6
lion	tack	536	619	-83	2.034	1.487	4.127	4.078	244	363	511	463	4	16
palm	jade	541	624	-83	2.364	1.850	4.025	4.043	333	572	515	359	6	7
hunter	outfit	544	516	28	2.796	1.608	4.007	3.989	342	417	428	489	4	2
reed	hose	722	703	19	2.555	1.620	3.989	3.970	369	314	430	449	18	12
theft	thief	552	479	73	2.775	1.986	3.970	3.970	386	322	499	529	1	2
rebel	waist	543	575	-32	2.815	1.954	3.912	3.932	461	325	448	540	3	3
arrival	tobacco	588	580	8	2.796	2.021	3.761	3.714	394	366	548	558	1	1
berry	fever	592	532	60	2.697	2.164	3.664	3.689	289	358	470	454	7	6
basin	charm	555	556	-1	2.918	2.076	3.638	3.638	250	456	504	514	4	6
hunger	gentry	509	673	-164	3.530	2.841	3.611	3.584	275	556	584	309	3	4
loyalty	upright	547	566	-19	2.920	2.028	3.497	3.526	497	436	491	480	3	2
envy	dent	547	649	-102	3.008	2.026	3.497	3.497	431	361	511	480	1	15
tyrant	kennel	710	620	90	4.161	3.122	3.296	3.296	492	322	387	449	1	4
linen	groan	573	598	-25	3.270	2.347	3.258	3.258	386	342	515	508	4	4
Mean					2.199	1.391	5.136	5.132	389	363	510	519	4.1	4.5

Note: H=High-CD, L=Low-CD, lnLF=log-transformed lexeme frequency, AoA=Age of Acquisition, N=Neighbourhood Density.

Nonword foils

phokus	ajerne	moovee	konvay	appruve
trubble	golph	leest	daylie	agreid
vidioh	famuss	abuv	ceit	pensill
mygrane	coam	bote	crapht	ressed
fraim	dants	nayvee	eyedea	elifant
knowtis	hokes	stuph	oddit	reech
cheet	voyce	baikur	baleef	burth
raung	publick	staireo	dred	feal
stewpid	younify	muphin	pannil	staup
frite	klawk	peese	carred	parsil
steem	sayle	souper	saifly	gloab
raize	pitrole	ekspect	taystee	conseed
fil	boi	tost	fome	stoan
undir	miks	sope	stryke	reezun
sawlid	diside	sute	teer	grean
whide	frenned	advurse	mawp	keap

Practice items

crane	furst	bunch	slite	beam
ment	amateur	bedrume	supper	blaque